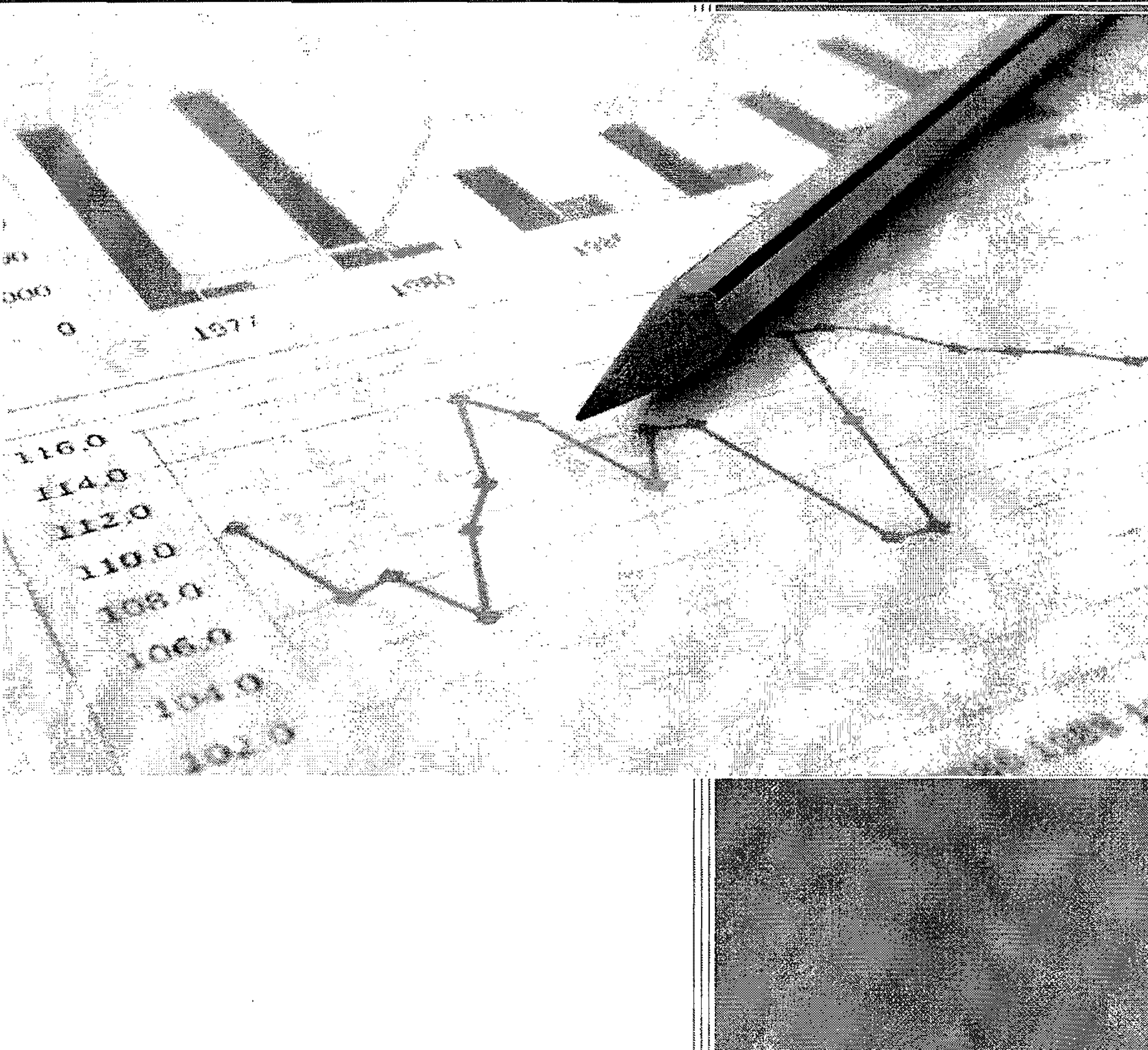
evolution

A Centre of Excellence for Civil Service Examination Guidance

B II, SECOND FLOOR, COMMERCIAL COMPLEX, BATRA CINEMA ROAD
NEXT TO ICICI A.T.M., DR. MUKHERJEE NAGAR, DELHI - 110 009
PHONE: 011 - 32974645, 32974651, 47092329

2017

Biostatistics



Content

Solution of Some Basic Decimal Problems, Logarithm and Antilogarithm

Basic Decimal problems :

Calculation of some basic decimal problems are given below for the convenience of Bioscience students who are supposed to be indifferent in mathematics.

Addition and subtraction. Numbers are kept in such a way that all decimal points lie just below each series and then addition or subtraction is done like simple addition or subtraction.

Example. Add 5.321, 0.012 and 37.9864.

$$\begin{array}{r} 5.321 \\ 0.012 \\ 37.9864 \\ \hline 43.3194 \\ \hline \end{array} \quad \textbf{Ans.}$$

Example. Subtract 7.7612 out of 31.614.

$$\begin{array}{r} 31.6140 \\ 7.7612 \\ \hline 23.8528 \\ \hline \end{array} \quad \textbf{Ans.}$$

Some tricky problems of subtraction :

Example. Subtract 8 out of zero (0).

Many students may say that the answer is 0 (zero) but answer is –8.

$$\begin{array}{r} 0 \\ -8 \\ \hline -8 \\ \hline \end{array} \quad \textbf{Ans.}$$

Example. Subtract – 8 out of zero (0).

At a glance it appears that answer is – 8 but answer is in positive number *i.e.*, +8

$$\begin{array}{r} 0 \\ -(-)\,8 \\ \hline +8 \\ \hline \end{array} \quad \textbf{Ans.}$$

Multiplication. Multiplication of decimal numbers is like simple multiplication. The only difference is to place the decimal in multiplicant after the total number of multiplier and multiplicator.

Example. 9.036×0.05 ; 0.000001×0.6 ; $.1 \times .1$.

$$\begin{array}{r} 9.036 \\ 0.05 \\ \hline 0.45180 \\ 0.0000\times \\ 0.000\times\times \\ \hline 0.45180 \\ \hline \end{array} \textbf{ Ans.} \qquad \begin{array}{r} 0.000001 \\ 0.6 \\ \hline .0000006 \\ \hline \end{array} \textbf{ Ans.} \qquad \begin{array}{r} 0.1 \\ 0.1 \\ \hline 0.01 \\ \hline \end{array} \textbf{ Ans.}$$

Fate of Positive and Negative sign during decimal multiplication follows the following rules :

$$(+)\times(+) = (+)$$
$$(+)\times(-) = (-)$$
$$(-)\times(-) = (+)$$
$$(-)\times(+) = (-)$$

Division. Following step is essential before solving the problem of division of decimal numbers.

— Right side of dividual and divider should be made equal digit. This is done by putting zero (0). Now decimal is removed and divide like normal division process.

Example. $0.102 \div 0.2$; $0.006 \div 500$.

0.2) 0.102 (

Now dividual have three digits after decimal while divider has only one. To make equal digits of both dividual and divider, we may put two zero (0, 0) after 2. Now we may remove decimal of both dividual and divider. Now divide in normal way.

(*i*) 200) 10200 (0.51 **Ans.**

$$\begin{array}{r} 1000 \\ \hline \times 200 \\ 200 \\ \hline \times \\ \hline \end{array}$$

Like wise (*ii*) 500) 0.006 (

Or

500000) 600000 (0.000012 **Ans.**

$$\begin{array}{r} 500000 \\ \hline 1000000 \\ 1000000 \end{array}$$

Problems of square root :

Symbol of square root is √. Square root of any number is the square of which is the root number.

Square root calculations can be done as follows :

— Pairing of sample number is done from right side. Odd number after decimal is paired by putting zero at right hand side.

— Now the following procedures are adopted to find square root shown in example.

Example. Find square root of 602.7025.

2	602.7025	24.55 **Ans.**
2	4	
44	202	
4	176	
485	2670	
5	2425	
4905	× 24525	
	24525	
	××××	

Basic rules to round up the decimal numbers :

— If after decimal last number is 5 or more than 5 then one number is added to left adjacent number. For example 3.966 can be written as 3.97.

— If last number after decimal is less than 5 then the same adjacent number is left. For example 3.964 can be written as 3.96.

— If last number after decimal is 5 and adjacent left number is odd then one number is added to it and written. If last number after decimal is even then the number is kept as it is. For example 3.965 is written as 3.96.

LOGARITHM

Literal meaning of logarithm is to short the calculations. The long process of addition, subtraction, multiplication and division is serially shortened by logarithm. For any positive number 'a' ($a \neq 1$), if $a^m = b$, then logarithm of 'b' to base 'a' is 'm' and it can be written as :

$\log_a^b =$, so we find that $a^m = b$ and $\log \frac{b}{a} = m$ are equal.

Integral power. For any real number 'a' and a positive integer n, we define a^n as follows :

$$a^n = a \times a \ldots\ldots \times a \text{ (n factor).}$$

If m and n are positive integers, and $m > n$, then

for $a \neq 0$

$$= \frac{a^m}{a^n} = \frac{a \times a \times \ldots\ldots\ldots \times a \ (m \text{ factors})}{a \times a \times \ldots\ldots\ldots \times a \ (n \text{ factors})}$$

$$= a \times a \times \ldots\ldots\ldots \times a \ \frac{(m-n) \text{ factor}}{1}$$

$$= a^{m-n}$$

If $m = n$, then $\frac{a^m}{a^m} = a^{m-m} = a^0$, *i.e.*, $a^0 = 1$.

Hence, by definition we take $a^0 = 1$.

Again, if we take $m = 0$ in $\frac{a^m}{a^n} = a^{m-n}$, we get $\frac{a^0}{a^n} = \frac{1}{a^n} = a^{0-n}$

Hence, for a positive integer n, we define $a^{-n} = \frac{1}{a^n}$.

We can now say that $\frac{a^m}{a^n} = a^{m-n}$, whether $m > n$

or $m = n$ or $m < n$.

a^n is called the nth power of a. The real number 'a' is called the base and 'n' is called the exponent of the nth power of 'a'.

Illustration :

(1) Write log form of $7^3 = 343$.

(2) Find log of 64 on base 4.

(3) Express in number of power of $\log 2^{128} = 7$.

(1) $7^3 = 343 = \log 7^{343} = 3$

(2) $a^x = N$, $4x = 64 = x = 3$ $[\because 4 \times 4 \times 4 = 64]$

or $4^3 = 4 \times 4 \times 4 = 64$

$\therefore$ log of 64 on base 4 is 3

or $\log 4^{64} = 3$

(3) $\log 2^{128} = 7$ or $2^7 = 128$

Simple logarithm and their uses :

Logarithms to base 10 : because the number 10 is the base of writing numbers, it is very convenient to use logarithms to the base 10. Some examples are as follows :

$\log 10^{10} = 1$ $(\because 10^1 = 10)$

$\log 10^{100} = 2$ $(\because 10^2 = 100)$

$\log 10^{1000} = 3$ $(\because 10^3 = 1000)$

$$\log 10^{10000} = 4 \qquad (\because\ 10^4 = 10000)$$
$$\log 10^{1} = 0 \qquad (\because\ 10^0 = 1)$$
$$\log 10^{0} = \alpha \qquad (\because\ 10^{\alpha} = 0)$$
$$\log 10^{1/10} = \log 10^{0.1} = -1 \qquad (\because\ 10^{-1} = 0.1)$$
$$\log 10^{1/100} = \log 10^{0.01} = -2 \qquad (\because\ 10^{-2} = 0.01)$$
$$\log 10^{1/1000} = \log 10^{0.001} = -3 \qquad (\because\ 10^{-3} = 0.001)$$

The above results indicate that if 'n' is an integral power of 10, *i.e.*, 1 followed by several zeros or 1 preceded by several zeros immediately to the right of the decimal point, the log n can be easily found.

If n is not an integral power of 10, then it is not easy to calculate log n. But mathematicians have made tables from which we can read off approximate value of the logarithm of any positive number between 1 to 10. And these are sufficient for us to calculate the logarithm of any number expressed in decimal form. For this purpose, we always express the given decimal as the product of an integral power of 10 and a number between 1 and 10.

Standard form of Decimal. Any number can be expressed in decimal form, as the product of (*i*) an integral power of 10 and (*ii*) a number between 1 and 10.

Example :

(*i*) 25.2 lies between 10 and 100

$$25.2 = \frac{25.2}{10} \times 10 = 2.52 \times 10^1.$$

(*ii*) 1038.4 lies between 1000 and 10000

$$\because \quad 1038.4 = \frac{1038.4}{1000} \times 10^3 = 1.0384 \times 10^3.$$

(*iii*) 0.005 lies between 0.001 and 0.01

$$\because \quad 0.005 = (0.005 \times 1000) \times 10^{-3} = 5.0 \times 10^{-3}.$$

In each case, we divide or multiply the decimal by a power of 10, to bring one non-zero digit to the left of the decimal point, and do the reverse operation by the same power of 10, indicated separately.

Thus any positive decimal can be written in the form

$$n = m \times 10^p$$

where 'p' is an integer (positive, zero or negative) and $1 \le m < 10$. This is called the "Standard form of n".

Working rule to use the logarithms table :

Following informations are essential to find the log value of any number :

(*i*) Log value is obtained by observing the round number.
(*ii*) Mantissa is never negative, it is always positive.

(*iii*) Move the decimal point to the left, or to the right, as per need, to bring one non-zero digit to the left of decimal point.

(*iv*) (*a*) If we move '*p*' places to the left, multiply by 10^p.

(*b*) If we move *p* places to the right, multiply by 10^{-p}.

(*c*) If we do not move the decimal point at all, multiply by 10^0.

(*d*) Write the new decimal obtained by the power of 10 (of step *ii*) to obtain the standard form of the given decimal.

Illustration :

1. Find out log of 45 (Two digit round number)
2. Find out log of 45.67 (decimal number)

1. First column of log table show 45. Second column below zero is 6532. Therefore, fraction of decimal of log 10^{45} is .6532. Because number consists of the two digits, therefore its characteristic will be 1.

Thus log $10^{45} = 1 + .6532 = 1.6532$.

In same fashion log value of three, four, five .... etc. number can be obtained.

2. Find out the log value of decimal numbers [such as (45.67)].

Following steps have to be taken to find out the log value of decimal numbers.

— Remove decimal to see the fraction of decimal of 4567.

— 45 lie in vertical column.

— For third number, 6, in horizontal column in 45 row, we find 5391. For fourth digit 7 mean difference is observed.

— In straight mean column below 7 is written 9. Add 5391 and 9.

— 5391 + 9 = 5400 is the fraction of decimal.

First of all write the characteristic and then put decimal and write fraction number of decimal. Thus log $10^{45.67}$.

$= 1 + 0.5400 = 1.5400$. **Ans.**

ANTILOGARITHM

Related corresponding number of a given log is called antilogarithm. To get the desired number we use the mantissa of logarithm and decimal point of that number is fixed by the characteristic of logarithm.

The first column of antilogarithm table consists of 0.00 to 0.99 and round up of four horizontal column is needed. For two digits after decimal is observed in vertical column. For third digit we move towards horizontal column in perpendicular to desired number. For fourth number we see in mean difference. If characteristic of logarithm (In case of positive number) add 1 and put decimal on left side as per obtained number. This is

antilogarithm value. If characteristic is negative then 1 is deducted from the characteristic and place the decimal after putting zero as many as number is obtained.

Note. Fraction of decimal of log must be positive while looking the antilogarithm. If it is not so; by mathematical manipulation negative value is converted to positive value and then antilogarithm table is consulted.

Example : Antilog of 3.8608 = 7244 + 13 = 7257 **Ans.**

Antilog of 2.1879 = 1538 + 3 = 154.1 **Ans.**

Antilog of $\bar{1}.2867$ = 1932 + 3 = .1935 **Ans.**

Antilog of $\bar{2}.2867$ = 1932 + 3 = .01935 **Ans.**

Antilog of $\bar{3}.2867$ = 1932 + 3 = .001935 **Ans.**

If fraction of decimal is negative then it is converted to positive by mathematical manipulations as follows :

Example. – 2.7133 whole number and mantissa both are negative. We can not use antilog table unless the number becomes positive.

$$\begin{aligned} -2.7133 &= -2 - 0.733 \\ &= -2 - 1 + 1 - 0.7133 \\ &= -3 + (1 - 0.7133) \\ &= -3 + 0.2867 = \bar{3}.2867 \end{aligned}$$

Now we may find the antilog value,

Antilog $\bar{3}.2867$ = 1932 + 3 = 0.001935 **Ans.**

Use of antilog in computing mathematical problems :

Example. Simplify $\dfrac{45 \times 20}{10000}$

Suppose $x = \dfrac{45 \times 20}{10000}$

Taking log both sides

$$\log x = \log . \frac{45 \times 20}{10000}$$

As per rule of log

$$\begin{aligned} \log x &= \log 45 + \log 20 - \log 10000 \\ &= 1.6532 + 1.3010 - 4 \\ &= 2.9542 - 4 = -1.0458 \\ &= -1 - 0.0458 \\ &= -1 - 1 + 1 - .0458 \\ &= -2 + (1 - 0.0458) \\ &= -2 + 0.9542 = \bar{2}.9542 \\ x &= \text{antilog of } \bar{2}.9542 \\ &= 8995 + 4 = 08999. \end{aligned}$$

Ans.

EXERCISE

1. Add 10.336 and 15.4417 [Ans. 25.7777]
2. Add 0.25 + 0.23 + 0.071 + 0.133 + 0.133 + 0.0625. [Ans. 0.9295]
3. Subtract – 0 from 0. [Ans. 0]
4. Subtract + 0 from + 0. [Ans. – 0]
5. Subtract 5 out of 0. [Ans. –5]
6. Multiply 36.5 with 0.25. [Ans. 8.825]
7. Divide 6.25 from 2.25. [Ans. 2.5]
8. Divide 2.5 from 0. [Ans. 0]
9. Divide 0 by 5.4. [Ans. 0]
10. Find out log of 45. [Ans. 1.6532]
11. Find out log of 45.67. [Ans. 1.5400]

1. Statistical Terms, Notations and Classification of Data

Synopsis : Introduction, Purpose and scope, statistical terms and notations, Collection and Presentation of data, classification of data, Qualitative and quantitative data. Preparation of Frequently distribution table. Preparation of cumulative frequency distribution table.

Introduction. Statistics is the *science of figures which deals with collection, analysis and interpretation of data.* Data is obtained by conducting a survey or an experimental study. The use of statistics in biology is known as Biostatistics or Biometry.

Purpose and scope of statistics. The purpose of statistics is not only to collect numerical data but is to provide a methodology for handling, analysing and drawing valid inferences from the data. It has wide application in almost all sciences—social as well as physical such as biology, psychology, education, economics, planning, business management, mathematics etc.

SOME IMPORTANT STATISTICAL TERMS AND NOTATIONS

STATISTICAL TERMS

While studying various aspects of problems of statistics one has to come across several statistical terms. Few important statistical terms are given below :

1. Population. The popular idea of population is universe. But statistician's idea of population is quite different from the popular idea. Biometric study regard the population of some limited region as its universe. *The population in a statistical investigation refers to any well defined group of individuals or of observations of a particular type.* In short one can say that *a group of study element* is called **population.** For example all fishes of one species present in a particular pond could be a population. All patients of a hospital suffering from AIDS may be considered as population while few patients are used as study elements.

2. Sample. In case of large population, it becomes practically impossible to collect data from all the members. In order to study the Haemoglobin percentage (Hb%) of patients of a hospital, it will be more convenient and quicker to collect data from few patients. Here patients taken for study are sample.

Sample may be defined as *fraction of a population drawn by using a suitable method so that it can be regarded as representative of the entire population.*

3. Variable. In everyday life, we come across living beings and phenomena, which vary in a number of ways, even though they belong to the same general category or type. Measurement of characteristics is called variable.

Animals of some species may differ in their length, weight, age, sex, Hb%, YO_2 intake, fecundity (Rate of reproduction), RBCs count, habits, personality traits etc. The above mentioned characteristics are variables. The variable may be defined as "*The characteristics on which individuals differ among themselves is called a variable.*" Variables may be of two types :

(*a*) **Quantitative variable.** Whenever the measurement of a characteristic is possible on a scale in some appropriate units, it is called a quantitative variable. Examples of quantitative variables are measurement of length, weight, age, intellectual ability etc. Quantitative variables can be further sub-divided into two types :

(*i*) Discrete or discontinuous variable and (*ii*) continuous variable.

(*i*) **Discrete or discontinuous variable** is one where the values of the variables differ from one another by definite amounts, *i.e.*, these vary only by finite '*jumps*' or '*breaks*'. For example the number of persons in a family or number of fish in a pond.

(*ii*) **Continuous variable** can assume all values within a certain interval and as such are divisible into smaller and smaller fractional units. Thus values of a continuous variable have no 'breaks' or 'jumps'. Measurement of length, weight, Hb%, VO_2 consumption, intelligence quotient (I.Q.) etc. are some examples of a continuous variable.

(*b*) **Qualitative variable.** It is unmeasurable variable and is unexpressible in magnitudes. But it can be expressed in quality. These qualities are called attributes. Colour of flower or animal, wrinkled seeds or smooth seeds etc. are examples of a qualitative variable.

4. Parameter. The numerical quantities which characterise a population (in respect of any variable) are called parameters of the population. For example, if the characteristic is length and measurements

of length is variable then the mean length can be regarded as a parameter. Usually all the important characteristics of a population can be specified in terms of a few parameters.

5. Statistics. Description of the properties of a population in terms of its parameters can be done with the help of statistical methods.

The term statistic is used to denote summary value of any quantity that is calculated from sample data. A statistic is usually calculated to provide an estimate of some population parameter. Thus the mean calculated from sample is a statistics that serves as an estimate of the parameter, population mean.

6. Observation. Measurement of an event is only possible by observation. For example Hb% in any animal is an event while 14 g/100 c.c., is a measurement, RBCs number is an event and 44 lacs/mm^3 is a measurement and these are observing experiments.

7. Data. A set of values recorded on one or more observational unit is called data. First step of statistical study is the collection of data. In scientific research work data is collected only from personal experimental study. Data collected by personal investigation is called primary data.

STATISTICAL STUDIES

In statistics one has to make use of certain formulae and computational procedures for the analysis of available data in order to describe its properties. In doing so during computational procedures certain statistical notations are universally used. Some common notations are mentioned as follows :

Summation. It is represented by the capital Greek letter Σ (Pronounced as sigma). (1) Σ stands for summation of observations, (2) Per cent—%, (3) Mean—$\bar{X}$, (4) Equal to : =, (5) Greater than—>, (6) Lesser than—<, (7) Observed number—O, (8) Expected number—E, (9) Degree of freedom—df, (10) Number of groups or classes—K, (11) Probability—p, (12) Deviation—x, (obtained from actual mean), (13) Deviation—x' (obtained from assumed mean), (14) Frequency—f, (15) Quartile deviation—Q, (16) Mean deviation—δ, (17) Standard deviation—σ, (18) Assumed mean—w, (19) Correction— c (20) Length of class-interval—i, (21) Quartile deviation—Q.

COLLECTION AND PRESENTATION OF DATA :

Collection of data. Statistical data is a set of facts expressed in quantitative form. The data can be obtained through primary source or secondary source. Data obtained by the investigator from personal

experimental study is called *primary data*. If the data is obtained from secondary source such as from Journals, Magazines, Paper, etc. it is known as *secondary data*. In scientific work only *primary data* are used.

Presentation of data. Data obtained by the investigator can be displayed in tabular form, diagrams and through charts. Display of data in tabular form is called classification of data and through charts is known as charting of data.

Process to arrange and present primary data in a systematic way is called classification of data. Data may be grouped or classified in following various ways :

(*i*) **Geographical**; *i.e.*, according to area or region. If we take into account production of fish or lac or silk statewise, this would be called geographical classification.

(*ii*) **Chronological**; *i.e.*, according to occurrence of an event in time. Egg production of a poultry farm for last five years are given below which is an example of chronological classification :

Year	*Egg production*
95—96	1590
96—97	1672
97—98	1882
98—99	1961
99—2000	2233

(*iii*) **Qualitative**; *i.e.*, according to attributes or quality. For example, if a species of fish in a pond is to be classified in respect to one attribute say sex, we can classify them into two groups. One is of males and other is of females.

When the classification is done with respect to one attribute, which is simple or dichotomous in nature, two classes are formed, one possessing the attribute and the other not possessing the attribute. This type of qualitative classification is called simple or dichotomous classification.

When we classify fishes simultaneously with respect to two attributes, *i.e.*, sex and infected condition, then fishes are first classified with respect to 'sex' into 'males' and 'females'. Each of these classes may then be further sub-divided into 'infected' and 'uninfected'. Thus the attribute sex and condition infection in fishes are classified into four classes, namely—(*a*) Male uninfected, (*b*) Male infected, (*c*) Female uninfected, (*d*) Female infected. The classification, where two or more attributes are considered and several classes are formed is called a manifold classification.

(*iv*) **Quantitative**; *i.e.*, according to magnitudes. For example, the thickness of a plant may be classified according to their growth rate. Quantitative data may be of two types :

(*a*) **Continuous data.** It covers all values of a variable. Hb% of a person can be expressed in any values such as 13 mg/100 c.c., 13.1 mg/100 c.c. and so on. Water percentage in the body of a species may be 65%, 65.1%, 65.2%, 65.3% and so on.

(*b*) **Discrete data.** The term discrete data is limited to discontinuous numerical values of a variable. It can be done only in whole number. For example number of persons in a family or number of books in a library can be said only in whole number. One can't say that there are $4\frac{1}{2}$ (Four and half) persons in my family or there are $500\frac{1}{2}$ books in this library.

PREPARATION OF FREQUENCY DISTRIBUTION TABLE

Quantitative data is grouped or classified and presented in the form of a frequency distribution table. The frequency distribution table presents the quantitative data very concisely indicating the number of repetition of observations. It records how frequently a variable occurs in a group study.

Following Raw data is obtained in an investigation. 100 pea plants bore pods ranging from 15 to 41 in a garden of pea plants (see Raw Data Table A) :

Raw Data Table A :

33,	31,	28,	15,	17,	17,	16,	18,	16,	18,	20,	22,	24,
25,	31,	27,	30,	29,	33,	28,	20,	22,	23,	25,	41,	39,
30,	36,	37,	27,	33,	28,	31,	29,	32,	31,	29,	34,	19,
22,	25,	40,	19,	21,	24,	30,	26,	37,	27,	28,	32,	32,
31,	29,	34,	21,	23,	25,	40,	26,	38,	27,	26,	33,	28,
34,	29,	30,	30,	35,	29,	23,	29,	26,	38,	27,	32,	28,
34,	35,	29,	30,	33,	32,	35,	29,	24,	26,	38,	27,	36
28,	34,	29,	35,	30,	33,	32,	36,	37,				

Raw Data Table B :

15,	16,	17,	17,	18,	18,	19,	19,	20,	20,	21,	21,	22,
22,	22,	23,	23,	23,	24,	24,	24,	25,	25,	25,	25,	26,
26,	26,	26,	26,	27,	27,	27,	27,	27,	27,	28,	28,	28,
28,	28,	28,	28,	29,	29,	29,	29,	29,	29,	29,	29,	29,
30,	30,	30,	30,	30,	30,	30,	31,	31,	31,	31,	31,	32,
32,	32,	32,	32,	32,	33,	33,	33,	33,	33,	33,	34,	34,
34,	34,	34,	35,	35,	35,	35,	35,	36,	36,	36,	37,	37,
37,	38,	38,	38,	39,	39,	40,	40,	41,				

Our first step in the preparation of frequency distribution table is to arrange them in ascending order of magnitude. The data is then said to be in array. The above raw data table A is arranged in ascending order of magnitude as shown in raw data table B.

Steps for the preparation of a discrete frequency distribution table may be taken as follows : (Table 1.1).

— A table of two columns is prepared. First column contains variables and second column contains repetition number of variable *i.e.*, frequency of variables.

— In above data variable 15 is obtained only once. Therefore frequency 1 is mentioned against variable 15. Variable 16 is obtained twice, therefore, frequency 2 is mentioned against this variable. In the same fashion frequencies of all variables of above data are mentioned and a frequency distribution table 1.1 is obtained.

Table 1.1

No. of Pods (Variables)	*No. of Plants (Frequency)*	*No. of Pods (Variables)*	*No. of Plants (Frequency)*
15	1	29	9
16	2	30	7
17	2	31	5
18	2	32	6
19	2	33	6
20	2	34	5
21	2	35	4
22	3	36	3
23	3	37	3
24	3	38	3
25	4	39	2
26	5	40	2
27	6	41	1
28	7		

For convenience discrete frequency table may be prepared with the help of tally mark. Following steps have to be taken to prepare discrete frequency table using tally mark :

— A table of three columns is prepared. In first column variables are mentioned. In second column repetition (frequency) of each variable is denoted by tally mark. In third column, total of tally mark, of each variable is written which is of course the frequency of variable.

— If variable appears only once then tally mark I is mentioned, for second repetition II, for third III, fourth IIII but for fifth a cut of fourth IIII is mentioned.

Following simple frequency table 1.2 is prepared using Raw data B in array with the help of tally mark.

Table 1.2

No. of Pods or Variable	*Tally mark*	*Repetition number of Plants of Frequency*
15	I	1
16	II	2
17	II	2
18	II	2
19	II	2
20	II	2
21	II	2
22	III	3
23	III	3
24	III	3
25	IIII	4
26	~~IIII~~	5
27	~~IIII~~ I	6
28	~~IIII~~ II	7
29	~~IIII~~ IIII	9
30	~~IIII~~ II	7
31	~~IIII~~	5
32	~~IIII~~ I	6
33	~~IIII~~ I	6
34	~~IIII~~	5
35	IIII	4
36	III	3
37	III	3
38	III	3
39	II	2
40	II	2
41	I	1

Preparation of frequency distribution table in class-intervals :

What is class interval and how it is prepared ?

To make data comprehensible one should classify or group identical values of the variables into ordered class intervals.

— To illustrate, the construction of a frequency distribution table in class interval, consider the raw data B, which represents the pods per plant in a garden.

Here we first decide about the number of classes into which data are to be grouped. Ordinarily, the number of classes should be between 5 and 20, but this may be done arbitrarily. The number of classes depends on the number of observations—with larger number of observations one can have more classes.

Now question arises about size of classes. The width or range of class is usually called *class-interval* and it is denoted by h. The width of *class-interval* must be of uniform size.

After deciding about *class-interval* we calculate range (the highest score H minus lowest score L or length of class interval) (H – L). From raw data B, Range of scores is R = 41 – 15 = 26 (Range is denoted by R). Now following formula may be applied to get the approximate number of classes which should expect to group the given observations.

$$\text{Number of classes} \quad k = \frac{\text{Range of scores}}{\text{Class interval}} = \frac{R}{h}.$$

Mid-point of class interval. Class mid-point is the sum of highest and lowest limits of class-interval divided by two. Thus the mid-point falls in the middle of upper and lower level of class-interval.

$$\text{Class mid-point} = \frac{\text{Highest limit of C.I.} + \text{Lowest limit of C.I.}}{2}$$

For example mid-point of a class interval 10—20 may be as follows :

$$\text{Mid-point of C.I.} = \frac{20+10}{2} = \frac{30}{2} = 15.$$

Frequency distribution table in class interval may be prepared in two ways :

(1) Overlapping and (2) Non-overlapping.

Overlapping frequency table. Values of variables are grouped in such a fashion that the upper limit of one class-interval is represented in next class interval. An *overlapping class interval frequency distribution table 1.3* may be prepared using data of table 1.2.

— Data of table 1.2 shows that the pods per plant ranges from 15 to 41. Pods *i.e.,* variables can be grouped into few class-intervals. Values of variables are grouped in such a fashion that the upper limit of one class interval is represented in next class interval.

— No. of pods ranges from 15 to 41. Range of first class interval may be 15—17, 17—19, 19—21 and so on. Plants having pods upto 17 minus one *i.e.,* 16 is kept in first class-interval (15—17). On perusal on Table 1.2 it appears that 3 plants come under this

group. Hence frequency of class interval 15—17 is 3. Plants having 17 no. of pods have to be included in next class interval (17—19). Four plants come under second class-interval (17—19). Hence frequency of class interval 17—19 is 4.

Non-overlapping class interval. Values of variables are grouped in such a fashion that the upper level of one class interval do not overlap the preceding class-interval. A non-overlapping frequency table 1.4 can be prepared using the data of table 1.2.

Here class interval may be 15—17, 18—20, 21—23, 24—26 and so on. Here upper level of one class interval do not overlap the lower level of next class interval.

Table 1.3. Overlapping Frequency Table.

No. of Pods in class interval	*No. of plants in frequency*
15—17	3
17—19	4
19—21	4
21—23	5
23—25	6
25—27	9
27—29	13
29—31	16
31—33	11
33—35	11
35—37	7
37—39	6
39—41	5
	$\Sigma f = 100$

Table 1.4. Non-overlapping Frequency Table.

No. of Pods in class interval	*No. of plants in frequency*
15—17	5
18—20	6
21—23	8
24—26	12
27—29	22
30—32	18
33—35	15
36—38	9
39—41	5
	$\Sigma f = 100$

On perusal of above two Tables (*Table 1.3—Overlapping frequency table and Table 1.4—Non-overlapping frequency table*) we find that the frequency of almost every class-interval differs from each other although original data of both is the same. The reason is placement of overlapping and non-overlapping scores in class-interval. In bio-science we usually use non-overlapping frequency table.

Preparation of cumulative frequency distribution table : Following steps have to be taken to prepare cumulative frequency table :

- A table of three columns is prepared.
- Cumulative frequency is determined by adding the frequency of a class interval with the frequency of the preceding class-interval.

Overlapping and non-overlapping cumulative frequency tables 1.5 and 1.6 are prepared using the data of tables 1.3 and 1.4 respectively.

Table 1.5. Overlapping cumulative frequency distribution table.

Class interval	*Frequency*	*Cumulative frequency*
15—17	3	3
17—19	4	3 + 4 = 7
19—21	4	7 + 4 = 11
21—23	5	11 + 5 = 16
23—25	6	16 + 6 = 22
25—27	9	22 + 9 = 31
27—29	13	31 + 13 = 44
29—31	16	44 + 16 = 60
31—33	11	60 + 11 = 71
33—35	11	71 + 11 = 82
35—37	7	82 + 7 = 89
37—39	6	89 + 6 = 95
39—41	5	95 + 5 = 100

Table 1.6. Non-overlapping cumulative frequency distribution table.

Class interval	*Frequency*	*Cumulative frequency*
15—17	5	5
18—20	6	5 + 6 = 11
21—23	8	11 + 8 = 19
24—26	12	19 + 12 = 31
27—29	22	31 + 22 = 53
30—32	18	53 + 18 = 71
33—35	15	71 + 15 = 86
36—38	9	86 + 9 = 95
39—41	5	95 + 5 = 100

Preparation of Non-overlapping frequency distribution table having different class intervals. Following three frequency distribution tables 1.7, 1.8, and 1.9 are prepared (using data of Table 1.1).

Table 1.7 having length of class interval 3, Table 1.8 having length of class interval 5 and Table 1.9 having length of class interval 10.

Table 1.7.

Class interval (No. of Pods) X	*Frequency (No. of plants) f*
15—17	5
18—20	6
21—23	8
24—26	12
27—29	22
30—32	18
33—35	15
36—38	9
39—41	5

Table 1.8.

Class interval (No. of Pods) X	*Frequency (No. of plants) f*
15—19	9
20—24	13
25—29	31
30—34	29
35—39	15
40—44	3

Table 1.9.

Class interval (No. of Pods) X	*Frequency (No. of plants) f*
15—24	22
25—34	60
35—44	18

Some basic rules in preparation of frequency distribution table :

— Size of class interval should not be too broad or too small. It should be preferably 3 to 10.

— The number of class intervals should not be too many or too few. 5 to 15 numbers of class interval is ideal.

— The class interval should be kept in ascending order for biological purposes.

— The heading of observation must be clear such as height in centimeters or inches, age in year or months, Hb in gm/100 mm etc.

— If certain data is omitted deliberately then the reason for doing so must be mentioned.

EXERCISE

1. Define, explain and mention uses of biometry.
2. What do you mean by data, population, sample, variable, parameter, class interval, frequency distribution, cumulative frequency distribution, primary data, secondary data.

3. What do you mean by grouping of data ? RBCs number in lac/mm^3 of 30 fishes of a species were measured as follows :
32, 30, 33, 31, 28, 33, 33, 32, 30, 32, 34, 31, 35, 35, 35, 35, 36, 36, 37, 34, 32, 34, 34, 36, 37, 38, 37, 36.

(*i*) Prepare simple frequency distribution table.
(*ii*) Prepare overlapping and non-overlapping frequency distribution table.
(*iii*) Prepare overlapping and non-overlapping cumulative frequency distribution table.

4. The lowest and highest levels of few class intervals are given below. Mention length and mid-points of each class interval :
40—50, 32—44, 20—32, 10—22, 20—30, 30—39, 40—49, 40—52, 53—64.

5. Explain following notations :

Σ, %, $\bar{X}$, x, x', Q, f, p, O, E,
σ, =, >, <, $\sim$, K, ||||.

6. The body weight (g) and Haemoglobin percentage (Hb%—g/100 ml) of 50 fishes of a species is given below. Make simple frequency table and non-overlapping cumulative frequency table :

$\frac{\text{Body weight (g)}}{\text{(Hob\%)}}$: $\frac{20.3}{8.3}$ $\frac{20.4}{8.4}$ $\frac{20.5}{8.5}$ $\frac{20.6}{8.6}$ $\frac{20.7}{8.7}$ $\frac{20.6}{8.4}$

$\frac{20.7}{8.5}$	$\frac{20.8}{8.6}$	$\frac{20.9}{8.7}$	$\frac{21}{8.8}$	$\frac{21}{8.6}$	$\frac{21.1}{8.7}$	$\frac{21.2}{8.8}$	$\frac{21.3}{8.9}$	$\frac{21.4}{9}$
$\frac{22}{8.9}$	$\frac{21.1}{9}$	$\frac{22.2}{9.1}$	$\frac{22.3}{9.2}$	$\frac{22.4}{9.3}$	$\frac{23.7}{11.5}$	$\frac{23.8}{11.6}$	$\frac{23.9}{11.7}$	$\frac{24}{11.8}$
$\frac{24.1}{11.9}$	$\frac{23.4}{12.4}$	$\frac{23.5}{12.5}$	$\frac{23.6}{12.6}$	$\frac{23.7}{12.7}$	$\frac{23.8}{12.8}$	$\frac{19.6}{11.6}$	$\frac{19.7}{11.7}$	$\frac{19.8}{11.8}$
$\frac{19.9}{11.9}$	$\frac{20}{12}$	$\frac{21.1}{9.5}$	$\frac{21.2}{9.6}$	$\frac{21.3}{9.7}$	$\frac{21.1}{9.8}$	$\frac{21.5}{9.9}$	$\frac{22.9}{9.4}$	$\frac{22}{9.5}$
$\frac{22.1}{9.6}$	$\frac{22.2}{9.7}$	$\frac{22.3}{9.8}$	$\frac{22.2}{12}$	$\frac{22.3}{12.1}$	$\frac{22.4}{12.2}$	$\frac{22.5}{12.3}$	$\frac{22.6}{12.4}$	

2. Graphical Representation of Biometric Data

Introduction. In previous chapter we have discussed that Raw data represented by frequency distribution makes the fact simple and easily understandable. After classwise or groupwise tabulation the frequencies of a characteristic can be presented by two kinds of drawings—*graphs* and *diagrams*. This makes the thing more easily understandable.

Some basic knowledge is essential for the preparation of *graph* of Statistical data. *Graph* is prepared with the help of two lines. The horizontal line of graph is called *Abscissa* or 'X' axis representing independent variable and the vertical line is known as *ordinate* or 'Y' axis representing dependent variable. The meeting point of 'X' and 'Y' axis is called zero (O) or origin point.

The right part of 'X' axis from the zero point ('0') is positive (+) and left part is negative (–). Likewise the upper part of 'Y' axis from zero point is positive while the lower part is negative. 'X' and 'Y' axis meet each other at 'O' point and graph divided into 4 parts. Each part is called Quadrant. Upper right part is called first Quadrant where 'X' and 'Y' both axis are positive. Upper left part is called 2nd Quadrant. Here 'X' axis is negative (–) and 'Y' axis is positive. The lower left part is called their quadrant where 'X' and 'Y' both axes are negative. The lower right part is known as Fourth Quadrant where 'X' axis positive (+) and 'Y' axis is negative (–). Mostly first quadrant is used for graphical representation of statistical data, where both axis are positive.

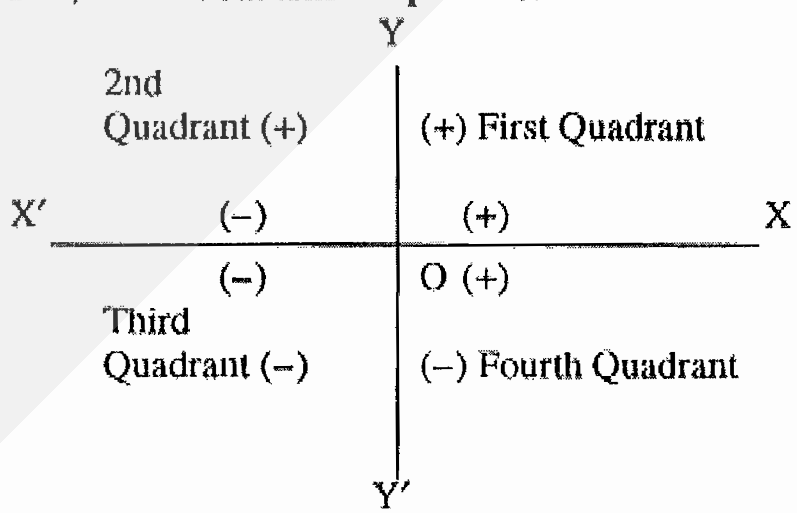


Fig. 2.1. Two axes "X" and "Y" meeting each other on "O" point producing 4 Quadrants.

Unit of representation. Small and appropriate unit bar line is required to represent the statistical data in graph. Suppose we have to represent large number such as 500, 1000, 1500, 2000 etc. on graph, then 1 cm on graph will represent 500 as follows :

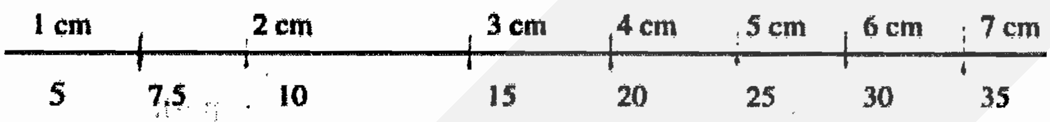
1 cm 2 cm 3 cm 4 cm 5 cm 6 cm 7 cm 8 cm 9 cm 9 cm 500 1000 1500 2000 2500 3000 3500 4000 4500

or in short

1 cm 2 cm 10 3 cm 15 4 cm 20 25 5 cm 6 cm 30 7 cm 35 8 cm 40 9 cm 45 cm 9 cm

1 cm 5 2 cm 7.5 3 cm 10 4 cm 20 5 cm 25 6 cm 30 7 cm 35 8 cm 40 9 cm 45

For number 750 on this line a point is mentioned between 5 and 10. The same method can be adopted to represent any number.



Presentation of *quantitative* and *continuous* data is done by graphs and those in common uses are :

1. Histogram
2. Frequency polygon
3. Frequency curve
4. Cumulative frequency curve or ogive
5. Scatter or dot diagram.

Presentation of *qualitative* and *discontinuous data* is done by diagrams and those in common use are :

1. Bar diagram.
2. Pie-chart or sector diagram.

Presentation of quantitative and continuous data :

1. Histogram. It is a graphical representation of frequency distribution. Variable is mentioned on the horizontal line (X axis *i.e.*, abscissa) while frequency is marked on the vertical line (Y axis *i.e.*, ordinate). Frequency of each group will form a column or rectangle. Such a diagram is called 'histogram'.

A histogram (Fig. 2.2) drawn on the basis of data of frequency Table 2.1.

Table 2.1

Class interval	*Frequency*
0—10	3
10—20	4
20—30	7
30—40	8
40—50	9
50—60	9
60—70	2
70—80	6
80—90	2
	N = 50

OX-axis 1 cm = 10 class interval which denotes the rate of reproduction.

OY-axis 1 cm = 1 frequency which represents the frequency of rate of reproduction.

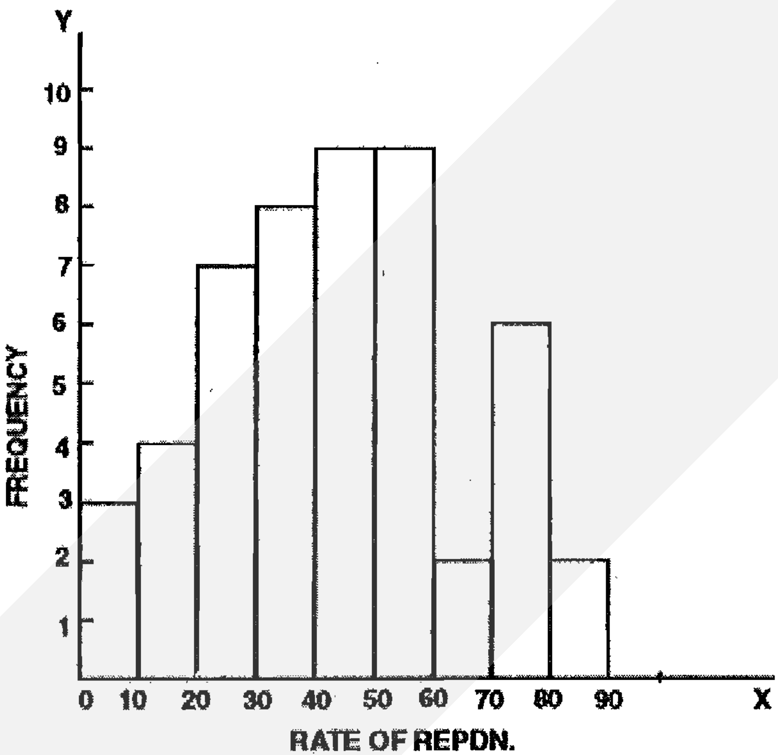


Fig. 2.2. Histogram showing rate of reproduction and frequency of 50 fishes of a species.

The frequency of class interval 0—10 is 3 which is being represented by 3 cm = 30 small square on OY axis (Because 10 small square = 1 frequency). In the same fashion rectangle for each class interval and frequency is prepared and a histogram (Fig. 2.2) is obtained.

2. Frequency polygon. It is also an area diagram of frequency distribution developed over a histogram. Join the mid-points of class intervals at the height of frequencies by straight lines. It gives a polygon, *i.e.*, a figure with many angles. (Following Fig. 2.3 is presented with the help of data of Table 2.1).

- Unbroken lines joining mid-points A, B, C, D, E, F, G, H and I of rectangle show the *frequency polygon*.
- Broken lines joining mid-points A to I of rectangle represent the *frequency curve*.

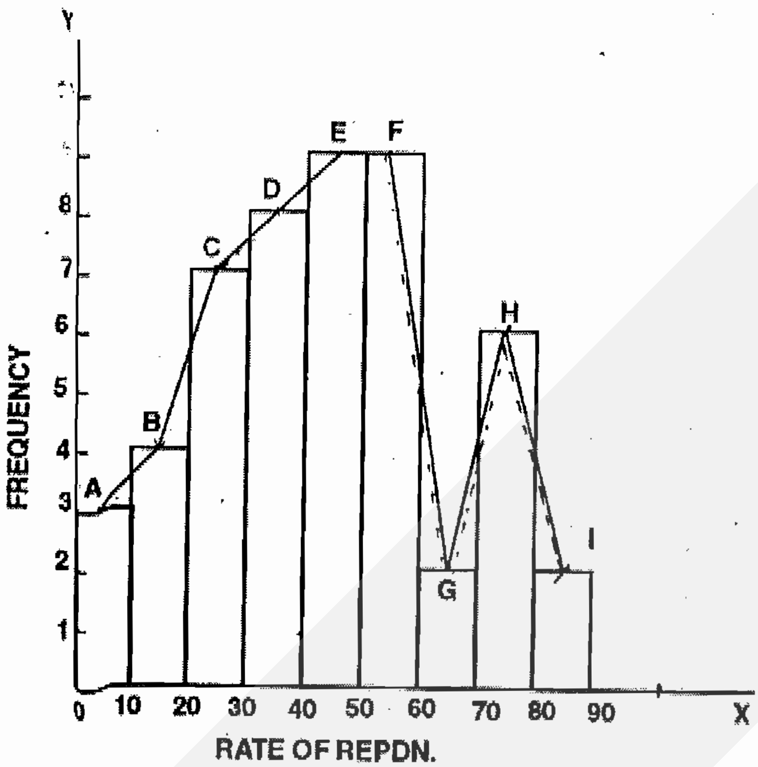


Fig. 2.3. *Frequency polygon* and *frequency curve* showing rate of reproduction in 50 fishes of a species and their frequency.

3. **Frequency curve.** When the number of observations is very large and group interval is reduced, the frequency polygon tends to lose its angulations giving place to a smooth curve known as frequency curves. This provides continuous graph giving the relative frequency for each value of an attribute. Broken lines of Fig. 2.3, represents the frequency curve, drawn by joining the mid-points of class-intervals of upper horizontal lines of rectangle by free hand. (Here the same data of Table 2.1 is used).

Relative frequency map can also be drawn with the help of relative frequency table. Relative frequency is the proportion of all observations determined by dividing the number of observations in a category by the total number of all observations. The relative frequency is generally calculated after the frequency distribution is obtained. The Table 2.2 given below suggests the method of calculation of relative frequencies.

Table 2.2.

Age of Albino Rats in months	*Number of Albino Rats in the category (Frequency)*	*Relative frequency Rf*
1—3	7	7 ÷ 43 = 0.16
4—6	8	8 ÷ 43 = 0.19
7—9	12	12 ÷ 43 = 0.28
10—12	10	10 ÷ 43 = 0.23
13—15	6	6 ÷ 43 = 0.14
	$\Sigma f = 43$	ΣRf = 1.00

Relative frequency (Rf) is generally calculated with the help of following formula : $Rf = \frac{f}{\Sigma f}$.

Relative frequency have been plotted in a graph whose relative frequency are shown in the vertical axis and variable (age groups) in the horizontal axis.

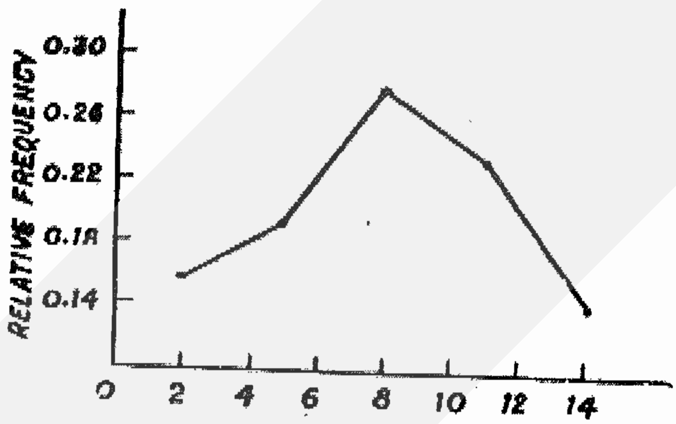


Fig. 2.4. Plotted value of Table 2.2, showing Relative Frequency.

4. Cumulative frequency curve or ogive. In previous chapter we studied that the cumulative frequency table is obtained by cumulating the frequency of previous classes. A cumulative frequency curve can be drawn with the help of following Table 2.3. Here calculation of cumulative frequencies relating to age group categories of 100 albino rats is shown.

Table 2.3.

Age group (X)	*No. of Albino Rats (f)*	*Cumulative Frequency (c.f)*
1—3	11	0 + 11 = 11
4—6	15	11 + 15 = 26
7—9	16	26 + 16 = 42
10—12	20	42 + 20 = 62
13—15	15	62 + 15 = 77
16—18	13	77 + 13 = 90
19—21	10	90 + 10 = 100

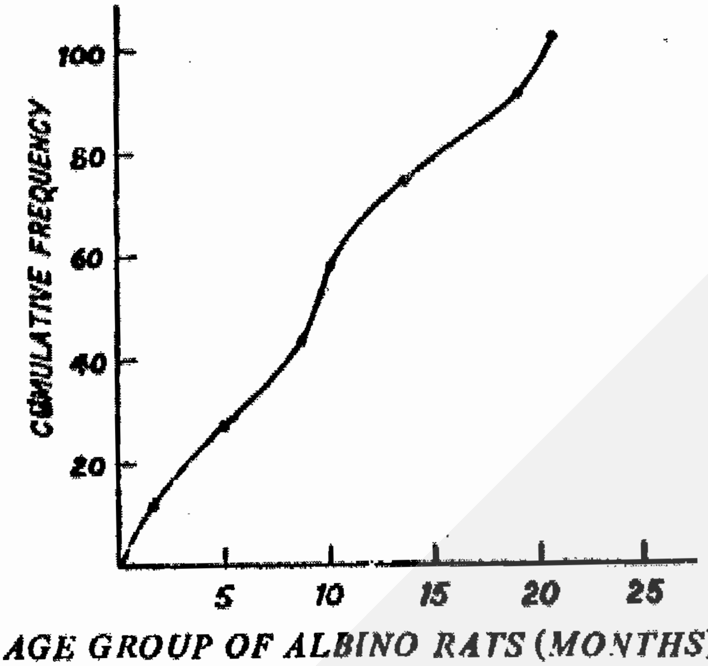


Fig. 2.5. Cumulative frequency curve or ogive.

The graphical display of cumulative frequency distribution as shown in Fig. 2.5, for age groups of albino rats shows one important feature, the curve is always in the ascending order. Such curves are called "ogives".

5. Scatter or dot diagram. It is prepared after cross tabulation in which frequencies of at least two variables have been cross classified. One variable being independent while the other dependent. An independent variable is the presumed cause of the dependent variable, In other words, it is a graphic representation, showing the nature of correlation between two variable characters X (independent) and Y (dependent) in the same persons or groups such as height and weight, Hb% and oxygen consumption, chronological age and reading age (growth age) etc.

The characters are read on the base (independent) and vertical (dependent) axes and the perpendiculars drawn from these readings meet to give one scatter point. When two axes are at right angles to each other, they are called orthogonal axes.

Length (independent) and *weight* (dependent) of 8 groups of fishes is given below to draw *scatter* or *dot diagram.*

Table 2.4.

Length of a species of fish	*Weight of the given fish*
13.9 cm	5.0 gms
15.7 cm	5.9 gms
15.8 cm	6.4 gms
17.5 cm	7.3 gms
18.1 cm	7.8 gms
19.9 cm	8.1 gms
22.0 cm	8.7 gms
23.8 cm	8.9 gms

Following mathematical custom, X, the independent variable, is the horizontal axis and Y, the dependent variable.

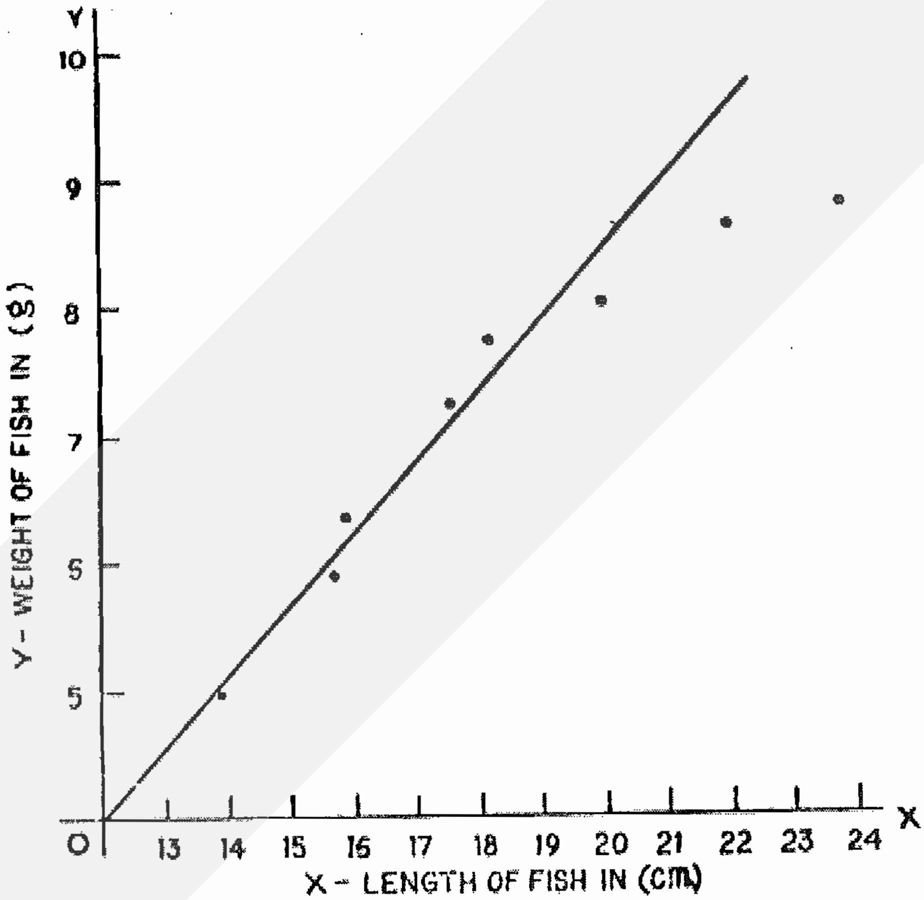


Fig. 2.6. Scatter diagram showing positive correlation between length and weight of 8 groups of fishes of a species.

Presentation of qualitative and discontinuous data :

(1) Bar diagram. Length of the bars drawn vertically or horizontally, indicates the frequency of a character. Bar diagram is an easy method

adopted for visual comparison of the magnitude of different frequencies in discrete data, such as mortality, immunisation status of population in different ages, sexes or places. Bars may be drawn in ascending or descending order of magnitude or in the series order of events. Spacing between any two bars should be nearly equal to half of the width of the bar. There are three types of bar diagrams :

(1) Simple bar diagram (Fig. 2.7).
(2) Multiple bar diagram (Fig. 2.8).
(3) Proportional bar diagram (Fig. 2.9).

Example : Data of oxygen consumption in different months of the year in a species of fish were obtained (between Jan. 77 to Jan. 78) and given below to draw a simple bar diagram (Fig. 2.7).

Table 2.5

Months	*Oxygen consumption c.c./kg/h*
Jan. 77	67
Feb. 77	74
Mar. 77	84
Apr. 77	85
May 11	100
June 77	105
July 77	95
Aug. 77	90
Sept. 77	90
Oct. 77	78
Nov. 77	74
Dec. 77	64
Jan. 78	62

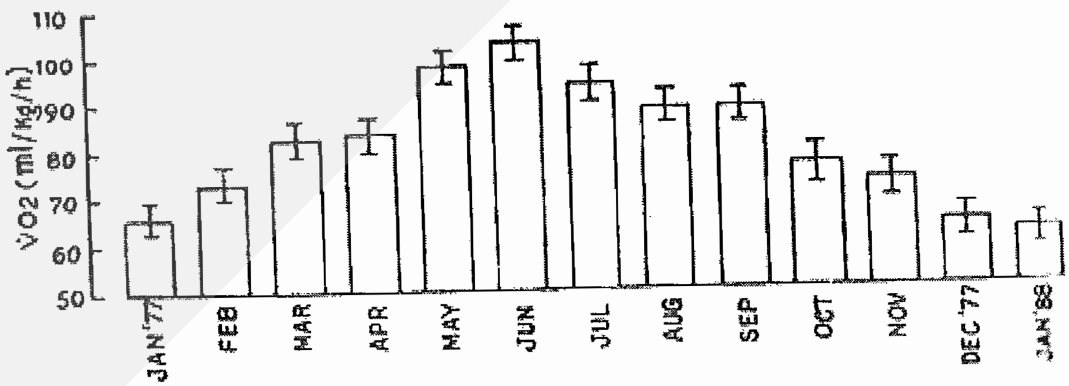


Fig. 2.7. Simple bar diagram showing oxygen consumption in a species of fish (M. aculeatus) in different months of a year.

Example : The mean values of different Haematological parameters (RBC's count, Hb%, PCV) of a species of fish was studied during different months of year (between Jan. 77 to Jan. 78) were obtained and are given below to draw a multiple bar diagram (Fig. 2.8).

Table 2.6

Months	*RBC (lac/mm³)*	*Hb (g/100 ml)*	*PCV (%)*
Jan. 77	2.01	8.5	14.1
Feb. 77	2.01	8.6	14.1
Mar. 77	2.08	8.8	14.1
Apr. 77	2.12	9.1	14.4
May 77	2.25	11.7	16.6
June 77	2.46	12.6	19.6
July 77	2.27	11.8	26.2
Aug. 77	1.87	9.7	24.9
Sept. 77	1.91	9.8	14.4
Oct. 77	2.30	12.2	14.5
Nov. 77	2.19	11.8	25.6
Dec. 77	2.12	10.9	24.2
Jan. 78	2.04	8.6	14.0

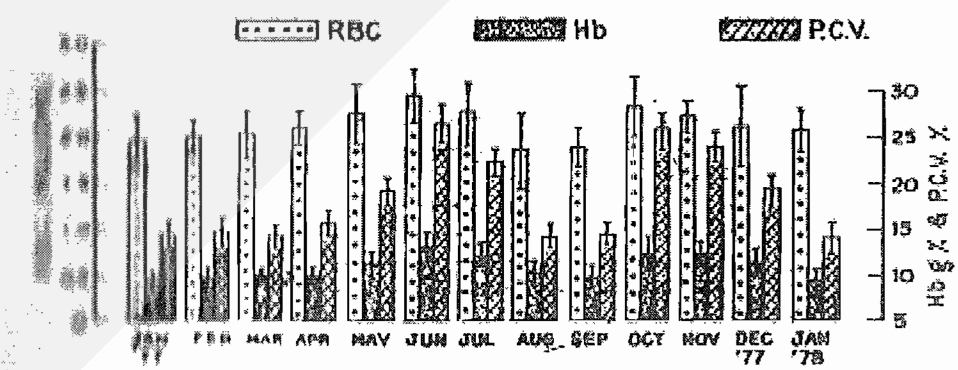


Fig. 2.8. Multiple bar diagram showing different Blood variables (RBC, Hb, PCV) in a species of fish (M. aculeatus) in different months of a year.

Monthly distribution of new and repeat patients in a hospital during different months of a year 1991 :

Table 2.7

Months	*Number of cases*		
	New	*Repeat*	*Total*
Jan. 91	2658	1114	3772
Feb. 91	2052	1470	3522
Mar. 91	2179	1610	3789
Apr. 91	1980	1351	3331
May 91	1714	1175	2889
June 91	2153	1434	3587
July 91	2203	1494	3697
Aug. 91	2123	1536	3659
Sept. 91	2642	1772	4414
Oct. 91	3055	1965	5020
Nov. 91	3869	2360	6229
Dec. 91	2557	1894	4451

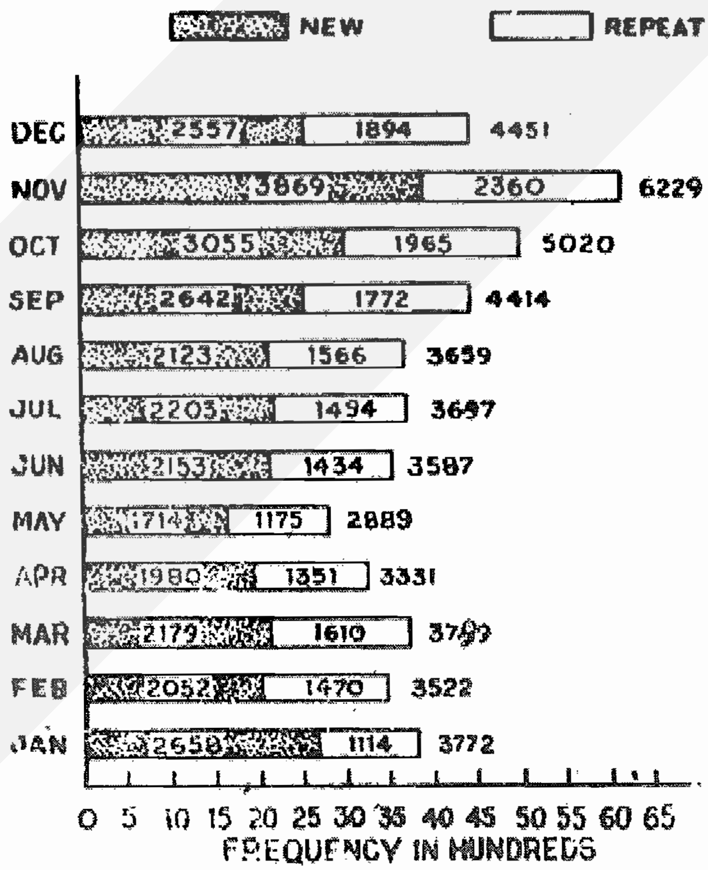


Fig. 2.9. Proportional bar diagram showing new and repeat outdoor patients given in Table 2.7

2. Pie Chart. (π chart) or sector diagram. It is another way of presenting discrete data of qualitative characters such as blood groups, Rh group, age groups, sex groups etc. The frequencies of the groups are shown in a circle. Degrees of angle denote the frequency and area of the sector. It presents comparative difference at a glance. Size of each angle is calculated by multiplying the class percentage with 3.6 *i.e.*, $\left(\frac{360}{100}\right)$ or by the formula $\frac{\text{Class frequency}}{\text{Total observation}} \times 360°$

Pie chart represent the data always in percentage.

Table 2.8. Distribution of blood groups in West Bengal Hindus.

Blood group	*No. of Persons*			*Percentage*	*Degrees*
	Male	*Female*	*Total*		
A	427	317	744	26.5	95.4
B	559	412	971	34.5	124.2
O	521	367	888	31.6	113.8
AB	122	85	207	7.4	26.6
Total	1629	1181	2810	100.0	360.0

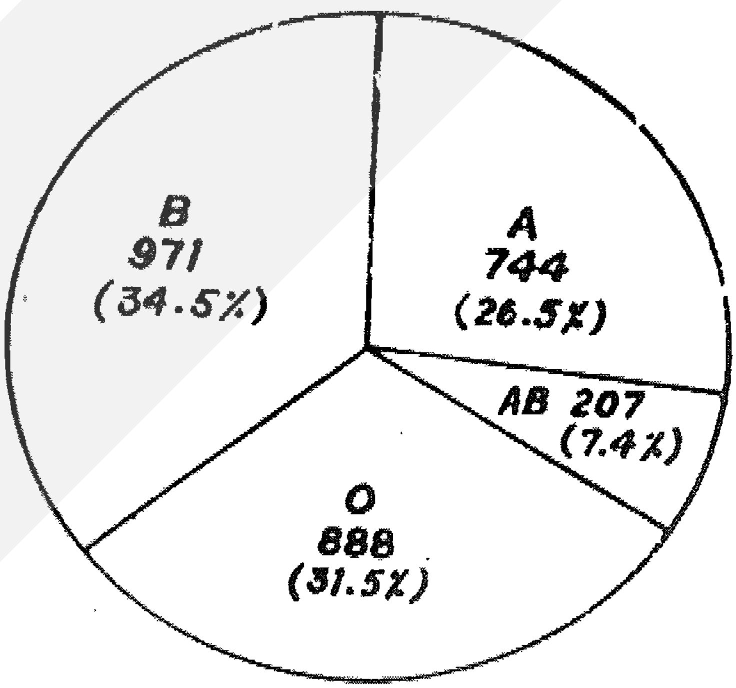


Fig. 2.10. Pie chart or sector diagram showing distribution of blood groups given in Table 2.8.

Size of angle for blood group A in Table 2.8 = 26.5 × 3.6 = 95.4

or $$\frac{744}{2810} \times 360 = 95.4.$$

Pie chart can be drawn showing distribution of blood groups as given in Table 2.8.

EXERCISE

1. What is abscissa and ordinate in a graph. Mention 4 quadrants giving a figure.
2. Draw histogram, frequency polygon, cumulative frequency curve with the help of following two tables :

Table I

Class interval	*Frequency*	*Class interval*	*Frequency*
1—10	3	51—60	40
11—20	14	61—70	47
21—30	21	71—80	50
31—40	25	81—90	50
41—50	40		

Table II

Class interval	*Frequency*	*Class interval*	*Frequency*
24.5—29.5	3	48.5—53.5	25
30.5—35.5	6	54.5—60.5	37
36.5—41.5	14	61.5—66.5	39
42.5—47.5	20	67.5—72.5	42

3. The percentage of water, lipid, protein and other materials are 66.35%, 6.66%, 5.2%, 21.79% respectively in the body of a species of fish. Draw a pie chart with the help of the given data.
4. Milk production of four states per day (litres) are as follows. Prepare a pie chart on the basis of following data :

States	*Bihar*	*Bengal*	*Delhi*	*U.P.*
Production (in 100 litres)	700	620	328	640

5. Rainfall in seven towns of Bihar were recorded as follows in the year 1997. Draw a pie chart on the basis of data :

Gaya	*Patna*	*Nawada*	*Ranchi*	*Arrah*	*Jehanabad*	*Sasaram*
150 cm	230 cm	125 cm	360 cm	240 cm	175 cm	210 cm

6. Following data were obtained in a hospital of Bombay in respect of age and frequency of cancer. Make a frequency polygon.

Age	39—49	50—59	60—69	70—79	80—89
No. of cancer patients	2	3	15	21	18

7. The weight of new born child (in kilogram) of Banaras Medical College in a particular day were as follows :
3.0, 3.5, 3.0, 2.5, 2.75, 2.0, 2.25, 2.25, 2.0, 2.2, 2.75, 2.0, 3.2, 3.5, 3.5, 4.0, 3.5, 3.2, 2.0, 2.5 and 2.0
Make cumulative frequency table and draw ogive.

8. Draw histogram with the help of following data :

No. of pods	*No. of plants*	*No. of pods*	*No. of plants*
15—17	5	30—32	18
18—20	6	33—35	15
21—23	8	36—38	9
24—26	12	39—41	5
27—29	22		

9. Draw a cumulative frequency diagram or ogive with the help of following frequency table :

Height of group in cm	*Frequency of each group*	*Cumulative class frequency*
160—162	10	10
162—164	15	25
164—166	17	42
166—168	19	61
168—170	20	81
170—172	26	107
172—174	29	136
174—176	30	166
176—178	22	188
178—180	12	200
	200	

3. Measures of Central Tendency

Synopsis : Measures of central tendency—Mean, Median and Mode.

Introduction. Central tendency may be considered as synonym of average. Average is a general term which describes the central value of a series, around which all other observations are dispersed.

There are two types of central tendency :

I. Mathematical Average and **II. Averages of Position**

I. Mathematical Average. Average represented mathematically is called mathematical average. There are three main types of mathematical averages :

(1) **Arithmetic mean,**
(2) **Geometric mean,**
(3) **Harmonic mean.**

II. Averages of Position. Average exhibited by position is called averages of position. There are two types of averages of position :

(1) **Median,** (2) Mode.

I. MATHEMATICAL AVERAGES :

(1) ARITHMETIC MEAN

Average obtained arithmetically is called Arithmetic mean. Arithmetic mean can be obtained both from ungrouped and grouped data :

Ungrouped data. Arithmetic mean is obtained by summing up all the observations and dividing it by the total number of observations.

Suppose each individual observation is $X_1, X_2, X_3, X_4, \ldots\ldots X_n$, sum of all observations in S, the number of observations is N.

Then $\Sigma X = X_1 + X_2 + X_3 + X_4 \ldots\ldots X_n$.

Thus arithmetic mean $(\bar{X})$ can be obtained with the help of following formula :

$$\bar{X} = \frac{\Sigma X}{N}$$

Here
$\bar{X}$ = Mean
Σ = Summation
X = Observation
N = Total number of observations.

Example 1. Haemoglobin percentage (Hb %) of 9 patients of a ward of hospital was obtained as 6 mg, 7 mg, 5 mg, 4 mg, 8 mg, 7 mg, 9 mg, 6 mg and 8 mg. Find out the arithmetic mean of the data.

Calculation :

$$\because \quad \overline{X} = \frac{\Sigma X}{N}$$

$$\therefore \quad \overline{X} = \frac{6+7+5+4+8+7+9+6+8}{9}$$

$$= \frac{60}{9} = 6.66 \text{ mg.} \quad \textbf{Ans.}$$

Example 2. WBC's number in 10 male frogs (*Rana tigrina*) are 8.19, 9.21, 10.40, 10.95, 12.14, 12.52, 13.41, 13.92, 14.78 and 15.74 lac/mm^3. Find mean WBC's number.

Calculation : For convenience a table is prepared as follows :

Table 3.1.

No. of observations	*Observations i.e. WBC's number*
1st frog	8.19
2nd frog	9.21
3rd frog	10.40
4th frog	10.95
5th frog	12.14
6th frog	12.52
7th frog	13.41
8th frog	13.92
9th frog	14.78
10th frog	15.74
N = 10	ΣX = 121.26

Here N = 10 and ΣX = 121.26

$$\overline{X} = \frac{\Sigma X}{N}$$

$$\overline{X} = \frac{121.26}{10} = 12.13. \quad \textbf{Ans.}$$

Grouped data. When data is presented in frequency distribution, then mean can be obtained by two methods ;

(1) Long method and (2) Short method ;

(1) Long method. Following formula is used to obtain mean by long method

$$\overline{X} = \frac{\Sigma fX}{\Sigma f}$$

Solved Example. Values of fecundity (Rate of reproduction) of 50 fishes of a species of fish (*Macrognathus aculeatus*) were obtained and on the basis of that, a frequency table is given below (Table 3.2). Calculate the mean value of fecundity by long method.

Table 3.2.

Class interval	*Mid point X*	*Frequency f*	*Multiplication of frequency and mid point f.X.*
1 – 10	$\frac{10+1}{2} = 5.5$	3	$5.5 \times 3 = 16.5$
11 – 20	$\frac{20+11}{2} = 15.5$	11	$15.5 \times 11 = 170.5$
21 – 30	$\frac{30+21}{2} = 25.5$	7	$25.5 \times 7 = 178.5$
31 – 40	$\frac{40+31}{2} = 35.5$	4	$35.5 \times 4 = 142.0$
41 – 50	$\frac{50+41}{2} = 45.5$	15	$45.5 \times 15 = 682.5$
51 – 60	$\frac{60+51}{2} = 55.5$	0	0
61 – 70	$\frac{70+61}{2} = 65.5$	7	$65.5 \times 7 = 458.5$
71 – 80	$\frac{80+71}{2} = 75.5$	3	$75.5 \times 3 = 226.5$
		$\Sigma f = 50$	$\Sigma fX = 1875$

Calculation : $\overline{X} = \frac{\Sigma f.X}{\Sigma f} = \frac{1875}{50} = 37.5.$ **Ans.**

Merits and demerits of arithmetic mean :

Merits : Arithmetic mean is the most important measures of central tendency because (*i*) It covers all the observations. (*ii*) It can be calculated easily and it expresses a simple relation between the whole and the parts. (*iii*) It does not get affected by the fluctuations of sampling. (*iv*) The mean of two or more series of observations can be had from the mean of the component series.

Demerits : (*i*) By observing data on graph; mean cannot be assumed. (*ii*) Mean obtained by calculation may not be represented by any series, (*iii*) By eliminating even a single series calculation becomes unreal.

(2) GEOMETRIC MEAN

This is the central tendency of a set of data following a Geometric progression. *In case of 'n' number of items the Geometric Mean is defined as the nth root of product of 'n' item of a observation or series.* Geometric Mean is denoted by G.M.

Computation of G.M. for ungrouped data :

Geometric Mean of items x_1 and x_2 will be $(x_1 \ldots\ldots x_n)^{1/2}$. Thus the equation for the Geometric Mean can be represented as

$$\sqrt[n]{x_1 . x_2 . x_3 \ldots\ldots x_n}$$

or $$(x_1 . x_2 . x_3 \ldots\ldots x_n)^{1/n}.$$

Example 1. On 1st March a baby weighted 14 lbs, on Ist May it weighted 20 lbs. What was the approximate weight of the said baby on 1st April.

Calculation : $GM = (x_1, x_2, x_3 \ldots\ldots x_n)^{1/n}$

or $GM = \sqrt[n]{x_1 . x_2 . x_3 \ldots\ldots x_n}$

Here ∵ $GM = (x_1 . x_2)^{1/2}$

$$= (14 \times 20)^{1/2}$$
$$= (2 \times 7 \times 2 \times 2 \times 5)^{1/2}$$
$$= 2(2 \times 7 \times 5)^{1/2}$$
$$= 2.\sqrt{2 \times 7 \times 5}$$
$$= 2\sqrt{70}$$
$$= 2 \times 8.36 = \mathbf{16.72}\ \textbf{Ans.}$$

The weight of the baby was **16.72** lbs, on 1st April.

Example 2. Find out Geometric Mean of 4, 32, 48 and 96.

Calculation : Here, according to the formula

$$GM = (4 \times 32 \times 48 \times 96)^{1/4}$$
$$= (2^2 \times 2^5 \times 2^4 \times 3 \times 2^5 \times 3)^{1/4}$$
$$= (2^{16} \times 3^2)^{1/4}$$
$$= 2^{16} \times \frac{1}{4} \times 3^{2 \times 1/4} \qquad [(a^m)^n] = a^{m \times n}$$
$$= 2^4 \times 3^{1/2}$$
$$= 16\sqrt{3}$$
$$= 16 \times 1.732 = 27.714.\ \textbf{Ans.}$$

Above problem may also be solved as follows :

$$GM = \sqrt[4]{4 \times 32 \times 48 \times 96}$$

$= \sqrt[4]{2\times2\times2\times2\times2\times2\times2\times2\times2\times2\times2\times3\times2\times2\times2\times3\times2\times2}$
$= \sqrt[4]{2^4\times2^4\times2^4\times3^2}$
$= 2\times2\times2\times2\sqrt{3}$
$= 16\sqrt{3} = 27.714$ **Ans.**

Computation of GM by logarithmic transformation for ungrouped data :

In an individual series of items the Geometric Mean is calculated by logarithmic transformation easily.

$$GM = \text{antilog of } \frac{\Sigma \log n}{n} \text{ or antilog of } \frac{(\log x_1 + \log x_2 \ldots.. + \log x_n)}{n}$$

Example. Find out the Geometric Mean of 2, 6, 9, 27 and 36.

Calculation : $\because GM = (x_1, x_2, x_3 \ldots\ldots x_n)^{1/n}$
$= (2\times6\times9\times27\times36)^{1/5}$
$= (2\times2\times3\times3^2\times3^3\times2^2\times3^2)^{1/5}$
$= (2^4\times3^8)^{1/5}$

Suppose GM $= x$
Then $x = (2^4\times3^8)^{1/5}$

Now both side log is used

$$\log x = \frac{1}{5}\{[\log(2^4 . 3^8)]\}$$

[Power of any number is converted into multiplication while taking log. As per rules log (A × B) = log A + log B, and log (A ÷ B) = log A – log B. Value of log is put with the help of log table. For example log 2 = 0.30103 and log 3 = 0.4771.]

$$\log x = \frac{1}{5}(\log 2^4 + \log 3^8)$$

$$= \frac{1}{5}(4 \log 2 + 8 \log 3)$$

$$= \frac{1}{5}(4\times0.30103 + 8\times0.4771)$$

$$= \frac{1}{5}(1.20412 + 3.81686)$$

$$= \frac{1}{5}(5.02108) = 1.004216.$$

Putting the value from antilog table.

GM or x = antilog × 1.00422 = 10.09 **Ans.**

Computation of GM for grouped data (discrete series). GM is the antilog of summed up values of products of frequency and log x.

Suppose frequency of scores X_1, X_2, X_3 ..... X_n are $f_1, f_2, f_3 .... f_n$ respectively.

$$\therefore \quad \text{GM}(g) = (X_1 . X_1 ... f_1 \text{ times}). (X_2 . X_2 ... f_2 \text{ times}) (X_3 . X_3 ... f_3 \text{ times}) (x_n . x_n .... f_n \text{ times})$$

$$= \left[x_1^{f_1} . x_2^{f_2} . x_3^{f_3} ..... x_n^{f_n}\right] \frac{1}{f_1 + f_2 + f_3 + f_n}$$

$$\therefore \quad \log(g) = \frac{1}{f_1 + f_2 + f_3 + f_n} \log \left| X_1^{f_1} . X_2^{f_2} . X_3^{f_3} ..... X_n^{f_n} \right|$$

Here $\quad n = (f_1 + f_2 + f_3 + ... + f_n)$

$$\therefore \quad \log(g) = \frac{1}{n} [f_i . \log x_1 + f_2 . \log x_2 + f_3 . \log x_3 ..... + f_n . \log x_n]$$

Here $\quad i = 1, 2, 3 ..... n$ (for integer)

$$= \frac{1}{n} \Sigma f_i \log x_i$$

$$g = \text{antilog} [\Sigma f_i . \log x_i]$$

Symbolically the GM = antilog $\dfrac{[\Sigma f . \log x]}{n}$.

Example. The number of Basophils (a kind of WBC) in blood of 30 patients of a hospital and their frequency were recorded as [Scores 11, 14, 17, 19, 22 and frequencies 5, 6, 8, 7, 4]. Find out the GM.

Calculation : Following table 3.3 was prepared on the basis of above data and log table.

Table 3.3.

Scores X	*Frequencies*	*log x*	*f. log x*
11	5	1.0414	5.2070
14	6	1.1461	6.8766
17	8	1.2304	9.8432
19	7	1.2788	8.9516
22	4	1.3424	0.3696
	$\Sigma f = 30$		$\Sigma f . \log x = 36.2480$

$$\text{GM} = \text{antilog} \left(\frac{\Sigma f . \log x}{\Sigma f}\right) = \text{antilog} \left(\frac{36.2480}{30}\right)$$

$$= \text{antilog} \quad (1.20826) = 16.15. \quad \textbf{Ans.}$$

Computation of GM from grouped data continuous series.

In continuous series the mid-point of various classes are taken for calculation.

The Geometric Mean, therefore, is the antilog of

$$\frac{\Sigma f.\ \log.\ \text{mid-pt}}{\Sigma f}.$$

Example. Calculate the GM from the following data :

No. of grains	*No. of panicles*
51—100	7
101—150	9
151—200	10
201—250	8
251—300	5

Calculations : Data is arranged in tabular form and mid-point of each class intervals is obtained and mentioned.

Table 3.4.

No. of grains or Class interval	*Mid-point X*	*No. of panicles f*	*log X*	*f log X*
51—100	75	7	1.88	13.16
101—150	125	9	2.10	18.90
151—200	175	10	2.24	22.40
201—250	225	8	2.35	18.8
251—300	275	5	2.44	12.2
		$\Sigma f = 39$		$\Sigma f.\ \log X = 85.46$

$$\text{GM} = \text{antilog}\ \frac{\Sigma f.\ \log X}{\Sigma f}$$

$$= \text{antilog}\ \left(\frac{85.46}{39}\right)$$

$$= \text{antilog of } 2.19 = 155.34. \quad \textbf{Ans.}$$

The geometric mean has the unique advantage that it is not affected by extreme values. More weightage is given to items of lower values than those of higher values.

(3) HARMONIC MEAN

Harmonic mean is reciprocals of arithmetic mean of the given observations.

Computation of harmonic mean for ungrouped data :

[The reciprocal of numbers in Arithmetic progression is called Harmonic progression. If Arithmetic progression is 2, 3, 4 ..... n then the harmonic progression is $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots \frac{1}{n}$.]

According to definition $x_1, x_2, x_3, \ldots\ldots x_n$, ($n$ numbers are in harmonic progression) then its reciprocal $\frac{1}{x_1}, \frac{1}{x_2}, \frac{1}{x_3}, \ldots\ldots \frac{1}{x_n}$ is its arithmetic progression. Then

$$\text{its arithmetic mean} = \frac{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\ldots\ldots+\frac{1}{x_n}}{n}$$

$$i.e., \quad \text{Arithmetic mean} = \frac{\Sigma X}{N}$$

According to definition :

$$\because \quad \text{Harmonic mean} = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\ldots\ldots\frac{1}{x_n}\right)}$$

$$\therefore \quad \frac{1}{HM} = \frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\ldots\ldots\frac{1}{x_n}\right).$$

Example. Haemoglobin percentage of five persons were measured as 1, 5, 10, 15 and 25. Find out the Harmonic Mean.

According to formula :

$$\frac{1}{HM} = \frac{1}{5}\left(\frac{1}{1}+\frac{1}{5}+\frac{1}{10}+\frac{1}{15}+\frac{1}{25}\right)$$

$$\frac{1}{HM} = \frac{1}{5}\left[\frac{150+30+15+10+6}{150}\right] = \frac{1}{5}\times\frac{211}{150} = \frac{211}{750}$$

$$\text{or,} \quad HM = \frac{750}{211} = 2.55. \quad \textbf{Ans.}$$

Grouped Data (Discrete series). To find out harmonic mean when data is grouped.

Suppose, frequencies of a score is $x_1, x_2, x_3, \ldots\ldots x_n$ is $f_1, f_2, f_3, \ldots\ldots f_n$ respectively.

Then
$$\frac{1}{H} = \frac{1}{\Sigma f}\left[\frac{f_1}{x_1} + \frac{f_2}{x_2} + \frac{f_3}{x_3} + \ldots\ldots + \frac{f_n}{x_n}\right]$$
$$= \frac{1}{\Sigma f} \ \Sigma\left(\frac{f}{x}\right)$$

Example. Hb % and its frequencies in 10 members of a family was studied and following results were obtained. Find out HM of the given series.

Hb% mg/100 ml.	*Frequencies*
12 mg	**3**
13 mg	**3**
14 mg	**1**
15 mg	**2**
16 mg	**1**

Following Table 3.5, is prepared to obtain HM.

Calculations :

Table 3.5.

Hb% (mg/100 ml) X	*Frequency* f	*l/X*	*f/X*
12	3	0.083	0.25
13	3	0.076	0.23
14	1	0.071	0.071
15	2	0.0666	0.133
16	1	0.0625	0.0625
	$\Sigma f = 10$		0.7465

$$\frac{1}{HM} = \frac{1}{\Sigma f} + \Sigma\left(\frac{f}{X}\right)$$
$$= \frac{1}{10} \times 0.7465$$
$$= \frac{0.7465}{10} = 07465$$
$$\therefore \qquad HM = \frac{1}{0.07465} = 13.39. \quad \textbf{Ans.}$$

II. AVERAGE OF POSITION :

1. MEDIAN

This type of average of Position (Median) indicate average position of a series. In a series all observations arranged in ascending or descending order and the middle observation is called the median.

The median is most suitable for expressing qualitative data such as colour, health, intelligence etc. Median is calculated differently for ungrouped and grouped data.

Ungrouped data. Median of ungrouped data is calculated by two different methods :

(*a*) When scores are in **odd number**. Formula to obtain Median is as follows :

$$\text{Median} = \left(\frac{n+1}{2}\right)\text{th item.}$$

Example. Hb% of an animal was recorded as 6, 7, 4, 5, 5, 3 and 4 gm/100 ml. Calculate the median.

Calculation : First of all above data is arranged in an ascending order — 3, 4, 5, 5, 6 and 7.

Total number of scores is 7 (an odd number)

$$\text{Mdn} = \left(\frac{n+1}{2}\right)\text{th item.}$$

$$= \left(\frac{7+1}{2}\right)\text{th item}$$

$$= \frac{8}{2}\text{th item} = 4\text{th item.}$$

Since 4th item of arranged data is 5, therefore

Median = 5 gms/100 ml. **Ans.**

(*b*) When the score is an **even number.** The total observations are added with 1 and divided by 2. The number thus obtained is the Median item. The Median falls between two observations. To get Median we have to add both observations and divide by 2. The above statement can be understood with the help of following example :

Example. RBCs. No. of 8 persons is 35, 44, 38, 36, 39, 40, 42 and 41 lac/mm^3. Find out median of this series.

Calculation : First of all data is processed in ascending order *i.e.*, 35, 36, 38, 39, 40, 41, 42 and 44 (lac/mm^3).

$$\text{Mdn.} = \left(\frac{n+1}{2}\right)\text{th number.}$$

Since the total observations are even number *i.e.*, 8.

$$\therefore \qquad \text{Mdn.} = \left(\frac{8+1}{2}\right)\text{th number}$$

$$= \frac{9}{2}\text{th number} = 4.5\text{th item.}$$

Here the Median is falling between the 4th and 5th observations denoted by 4.5. The 4th observation in arranged data is 39 and 5th observation is 40.

Therefore, Median between 4th and 5th number is as follows :

$$\text{Median} = \frac{39(4\text{th item}) + 40(5\text{th item})}{2}$$

$$= \frac{79}{2} = 39.5. \quad \textbf{Ans.}$$

In biological experiments the data may not be as simple as given in above examples, mostly the data is in the form of a grouped series.

Grouped data. Following formula is used to calculate Median from grouped series :

$$\text{Median} = l_1 + \frac{\left(\frac{N}{2} - F\right) \times i}{fm}$$

Here l_1 = The lower limit of that class interval where Median falls.

fm = The frequency of that class interval where Median falls

F = The cumulative frequency just above that class interval where Median falls.

i = The width of the class interval.

Example. Find the median from the following data which shows the fecundity of a species of fish :

Fecundity in C.I *X*	*Frequency* *f*
1—10	3
11—20	15
21—30	2
31—40	8
41—50	11
51—60	4
61—70	1
71—80	6

Calculation : Following cumulative frequency table is prepared using above data :

Table 3.6.

Class interval	*Frequency No. of fishes*	*Cumulative frequency*
1—10	3	3
11—20	15	3 + 15 = 18
21—30	2	18 + 2 = 20
31—40	8 Mdn.	20 + 8 = 28
41—50	11	28 + 11 = 39
51—60	4	39 + 4 = 43
61—70	1	43 + 1 = 44
71—80	6	44 + 6 = 50

Here, $\Sigma f = 50$

$\because$ Mdn. will fall in $\left(\frac{N+1}{2}\right)$th item.

$\therefore$ Mdn. $= \left[\frac{50+1}{2}\right]$th item.

$= \left[\frac{51}{2}\right]$th item $= 25.5$th item.

This 25.5th item lies in between 31—40 class interval.

Therefore 31—40 is the class interval where median falls. The lower limit of this class interval is 31.

Thus, $l_1 = 31$, $\Sigma = 50$, $F = 20$, $fm = 8$ and $i = +10$.

$$\text{Median} = l_1 + \frac{\left[\frac{\Sigma f}{2} - F\right] \times i}{fm}$$

$$= 31 + \frac{\left[\frac{50}{2} - 20\right] \times 10}{8}$$

$$= 31 + \frac{(25 - 20) \times 10}{8}$$

$$= 31 + \frac{5}{8} \times 10$$

$$= 31 + \frac{50}{8}$$

$$= 31 + 6.25 = 37.25. \quad \textbf{Ans.}$$

Computation of Median in Grouped data discrete series.

Median from grouped data having no class interval may be calculated using following formula :

$$\text{Median} = l_1 + \frac{(m - F)}{fm} \times (l_2 - l_1).$$

Here, l_1 = score value of median class

l_2 = score value above median class

m = median item

F = cumulative frequency above median item

fm = frequency of median item.

Example. Percentage of body water of 15 fishes and their frequency is given as follows. Find median of the given data :

Water %	60	62	64	70	72	74	76	78	82	84	86
Frequency	1	1	1	2	1	1	2	1	1	3	1

Cumulative frequency table 3.7 is prepared using the data of example.

Calculations :

Table 3.7.

Water %	*Frequency*	*Cumulative frequency*
60	1	1
62	1	1 + 1 = 2
64	1	2 + 1 = 3
70	2	3 + 2 = 5
72	1	5 + 1 = 6
74	1	6 + 1 = 7
76	2 Mdn.	7 + 2 = 9
78	1	9 + 1 = 10
82	1	10 + 1 = 11
84	3	11 + 3 = 14
86	1	14 + 1 = 15

Median will fall in $\left[\frac{N+1}{2}\right]$th item.

$$= \frac{15+1}{2} \text{th item} = \frac{16}{2} = 8\text{th item}.$$

Here : ∵ Score of 8th item = 76

∴ l_1 = 76

N = 15

F = 7

$$fm = 2$$
$$l_2 = 78$$
$$m = 8$$

$$\text{Median} = 76 + \left\{\frac{8-7}{2}\right\} \times \{(78-76)\}$$

$$= 76 + \left\{\frac{1}{2} \times (78-76)\right\} = 76 + \left\{\frac{1}{2} \times 2\right\}$$

$$= 76 + 1 = 77. \quad \textbf{Ans.}$$

Merits and demerits of median :

Merits :

(*i*) Median is a better indicator of an average than mean when one or more of the lowest or the highest observations are wide apart or not so evenly distributed.

(*ii*) It is calculated easily and can be exactly located.

(*iii*) The value of median is not influenced by abnormally large or small values or the change of any one value of the series.

(*iv*) It can also be used in qualitative measures.

Demerits :

(*i*) Arithmetic explanation of median is not possible.

(*ii*) To obtain median, data must be kept in ascending or descending order.

(*iii*) It gives equal importance to all series.

2. MODE

Mode is that value which is repeated maximum times in a series. In other words we can say that the mode is that value which has the maximum frequency.

Mode can be obtained by two methods :

Determination of mode at a glance. The value which is repeated maximum times in a series is considered as mode.

Example. Water percentage of fifteen fishes of a species of fish were recorded as 60, 64, 62, 76, 70, 74, 70, 84, 82, 72, 76, 84, 78, 84 and 86. Find the mode of this series.

Solution. First of all, data is arranged in ascending order. Not even single observation is spared. It comes as 60, 62, 64, 70, 70, 72, 74, 76, 76, 78, 82, 84, 84, 84 and 86.

By simple observation one can say that 84 is the mode, because 84 is repeated maximum times (three times) in the above series.

The above data can be arranged in following tabular form to ascertain the mode easily.

Table 3.8.

Water %	*Repetition*	*Water %*	*Repetition*
60	1	76	2
62	1	78	1
64	1	82	1
70	2	84	3
72	1	86	1
74	1		

Inference. On perusal of the table we find that repetition of 84 is maximum times *i.e.*, three times, none of other items appear so frequently. Therefore 84 is the mode of the series.

Detection of mode by method of grouping. In an experimental data if many items possess maximum or same frequency then we find out mode by method of grouping.

Example. Ovary weight and their frequency in a species of 50 fishes were recorded as follows :

Ovary weight (*g*) —	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Frequency —	**4**	**2**	**6**	**4**	**9**	**9**	**7**	**5**	**1**	**3**

Find out the mode of this series.

Solution. From the above data we find that item 25 and 26 both have maximum and same frequency *i.e.*, 9. Now question arises which item should be considered as mode.

To solve this problem, method of grouping is used to obtain mode. With the help of grouping method following Table 3.9 is prepared.

Table 3.9.

Wt. of ovary	*Frequency*					
	Each	*Two's*		*Three's*		
	I	*II*	*III*	*IV*	*V*	*VI*
21	4					
22	2	6				
23	6		8	12	12	
24	4	10				
25	9		13			19
26	9	18		22	25	
27	7		16			
28	5	12		13		21
29	1		6		9	
30	3	4				

Above table reveal how grouping has been done. In column I the frequencies have been given. In column II and III the frequencies (shown in column I) are grouped in two's. In column II starting from the first cumulative frequency of Ist and 2nd, 3rd and 4th, 5th and 6th, 7th, 8th, 9th and 10th have been shown.

In column III, the frequencies have again been paired and the total shown, but with a difference, the pairing is started from second item (instead of first item as in column II). As a result, the last item cannot be paired and has been left out. In column IV the grouping is in three's starting from the first item and the frequencies totalled for each group are shown. In this case also the last item cannot be grouped with any other one and has been ignored. In column V, the grouping is in three's, but the first item has not been considered and the grouping has started from the second item. In column VI also the grouping is in three's but first two item have been ignored. Here the last two items are left without a third partner and have not been considered. The total of the frequencies of each group are shown.

The analysis of the above data is done as follows :

Table 3.10.

Columns	*Size of item having maximum frequency*				
I		25,	26		
II		25,	26		
III			26,	27	
IV	24,	25,	26,		
V		25,	26,	27,	
VI			26,	27,	28
Total	1	4	6	3	1

On perusal of the above table it appears that item No. 26 appeared maximum times *i.e.*, six times. Therefore mode = 26.

Mode obtained with the help of formula. Three formulae are used to calculate mode from grouped series :

Formula No. 1 :

$$\text{Mode} = \frac{\begin{array}{c}\text{Maximum level of C.I. where}\\ \text{maximum frequency falls}\end{array} + \begin{array}{c}\text{Lowest level of C.I. where}\\ \text{maximum frequency falls}\end{array}}{2}$$

Formula No. 2 : Mode = 3 Mdn. – 2 Mean.

Formula No. 3 : $l + \left[\frac{(fm_1)}{fm_1 + fm_2}\right] \times i$

Here l = Actual lower level of that C.I. where maximum frequency falls.

fm_1 = frequency of just above of that C.I. where maximum frequency falls.

fm_2 = frequency of just below of that C.I where maximum frequency falls.

i = length of class interval.

Example. RBCs No. of 15 patients of a ward were recorded as 30, 32, 31, 38, 35, 37, 35, 42, 41, 36, 38, 42, 39, 40 and 44 lac/mm^3.

Find out modes of above observations using all the three formulae.

Sol. First of all a frequency table is prepared as follows :

Table 3.11.

C.I.	*f.*
30—34	3
35—39	7
40—44	5
	$f = 15$

According to Formula No. 1.

$$\text{Mode} = \frac{\begin{array}{c}\text{Maximum level of C.I. where} \\ \text{maximum frequency falls}\end{array} + \begin{array}{c}\text{Lowest level of C.I. where} \\ \text{maximum frequency falls}\end{array}}{2}$$

In the given Table 3.11 maximum frequency falls in class interval 35—39.

$$\text{Therefore mode} = \frac{39+35}{2} = \frac{74}{2} = 37 \quad \textbf{Ans.}$$

(Mode obtained by above formula is not considered to be very authentic therefore another two formula are used.)

Formula No. 2. Mode = 3, Mdn. – 2 Mean.

Example. Data of table 3.11 is used as example. Following table 3.12 of four columns is prepared to obtain mode with the help of above formula.

Sol. **Table 3.12**

Class-interval	*Mid-point*	*Frequency*	*Frequency × Mid-point*
30—34	32	3	32 × 2 = 96
35—39	37	7	37 × 7 = 259
40—44	42	5	42 × 5 = 210
		$\Sigma f = 15$	$\Sigma f.X$ = 565

For above formula one has to calculate mean and median.

$$\text{Mean} = \frac{\Sigma fX}{\Sigma f} = \frac{565}{15} = 37.4$$

$$\text{Mdn.} = l + \frac{\left(\frac{\Sigma f}{2} - F\right)}{fm} \times 5$$

$$= 34.5 + \frac{\left(\frac{15}{2} - 3\right)}{7} \times 5$$

$$= 34.5 + \frac{(7.5 - 3)}{7} \times 5$$

$$= 34.5 + \frac{4.5}{7} \times 5$$

$$= 34.5 + (.64) \times 5$$

$$= 34.5 + 3.20 = 37.7.$$

Putting the value of mean and median in above formula

$$\begin{aligned} \text{Mode} &= 3\ \text{Mdn.} - 2\ \text{Mean} \\ &= 3 \times 37.7 - 2 \times 37.4 \\ &= 113.10 - 74.8 = 38.3\ \textbf{Ans.} \end{aligned}$$

Formula No. 3 :

$$\text{Mode} = l + \left\{\frac{(fm_1)}{(fm_1 + fm_2)}\right\} \times i$$

Example. Taking data of table 3.12 into account $l = 34.5$, $f = 7$, $fm_1 = 3$, $fm_2 = 5$, and $i = 5$.

Sol. Therefore mode $= 34.5 + \left[\frac{(3)}{(3+5)}\right] \times 5$

$$= 34.5 + (4.375)$$

$$= 34.5 + 1.875 = 36.37.\ \textbf{Ans.}$$

Merits of mode :

(*i*) Mode is the quickest way to measure the central tendency.

(*ii*) It represents that series where maximum concentration of observations are present and therefore extreme items do not affect it.

Demerits :

(*i*) Mode is rarely used for medical and higher biological scientific calculations because calculation is not based on all series of data.

(*ii*) Arithmetic explanation of mode is not possible.

(*iii*) Sometimes it is indefinite.

(*iv*) It becomes difficult to calculate when many series of maximum frequency is present in data.

EXERCISE

1. Define and explain mean, median and mode. Mention formula to find out mean, median and mode both for ungrouped and grouped data.

2. Calculate mean and median from the following ungrouped data :

   (*i*) 10, 8, 20, 22, 39 and 18. [Ans. $\overline{X}$ = 19.34, median = 19]

   (*ii*) 19, 21, 17, 16, 19, 21, 23 and 23. [Ans. $\overline{X}$ = 19.8, median = 20]

   (*iii*) 17, 19, 13, 17, 13, 11 and 21. [Ans. $\overline{X}$ = 15.8, median = 17]

   (*iv*) 15.8, 13.3, 15.2, 13.3, 17.8, 18.1 and 18.9. [Ans. $\overline{X}$ = 16.05, median = 15.8]

   (*v*) 1060, 1060, 1070, 1130, 1370, 1190, 1270 and 1170. [Ans. $\overline{X}$ = 1165, median = 1150]

   (*vi*) No. of animals per cage
   3, 7, 8, 11, 13, 15, 16, 17, 18 and 20. [Ans. $\overline{X}$ = 12.8]

   (*vii*) Haemoglobin percent in g/100 ml
   6.0, 6.5, 7.5, 8.2, 8.5, 8.7, 8.8, 8.9, 9 and 9.5 [Ans. $\overline{X}$ = 8.16]

3. Compute N (no. of observation) when $\overline{X}$ (mean) = 5 and ΣX = 30.

   [Ans. 6]

4. The following data was obtained in a grass land community. Calculate mean, median and mode by ungrouped and grouped series.

   Number of seeds per plant (Indigofera species) :

   39, 55, 35, 45, 49, 52, 48, 33, 48, 47, 50, 51, 53, 50, 55, 53, 54, 53, 50, 48, 49, 31, 33, 50, 55, 50, 51, 50, 53, 55, 52, 49, 51, 50, 50, 44, 51, 50, 58, 59, 57, 59, 60, 58, 51.

   [Ans. Mean = 49.87, Median = 52.40]

5. Calculate mean, median and mode from the data given in following three tables :

Table A

Class interval	*Frequency*
16—20	4
21—25	4
26—30	9
31—35	7
36—40	13
41—45	3
46—50	3
51—55	2
56—60	2
61—65	3

Table B

Class interval	*Frequency*
30—39	5
40—49	4
50—59	3
60—69	8
70—79	9
80—89	10
90—99	4
100—109	3
110—119	2
120—129	2

Table C

Class interval	*Frequency*
50—54	2
55—59	1
60—64	3
65—69	1
70—74	7
75—79	31
80—84	6
85—89	7
90—94	1
95—99	3

Ans.

Table A
Mean = 77.24
Median = 36.38
Mode = 38.00

Table B
Mean = 66.3
Median = 75.5
Mode = 84.5

Table C
Mean = 77.09
Median = 77.74
Mode = 77

4. Measures of Dispersion (or Variation)

Synopsis* : *Introduction, Definition of Dispersion or Variation, Different measures of variation—The Range, Quartile Deviation (or Semi-inter Quartile range), Mean deviation, Standard deviation and Variance.

Introduction. The measure of central tendency is a single representative value for a set of data. The central value alone cannot give the total picture of a set of data. Therefore we use a method called Measure of Variation which may also be called as dispersion or deviation.

The measure of dispersion is one of the important tools of statistics for biologists because biological phenomena are more variable than that of physical and chemical sciences. For example Hb%, RBCs number, VO_2 consumption, Fecundity etc. of two individuals of same species of same age, sex, weight etc. will differ definitely. Cure rate with the same drug varies in different patients of same age and sex. Hb%, RBCs number, VO_2 consumption. Fecundity etc. of same individual will differ in different physiological (endogenous) or exogenous conditions.

Definition. A measure of variation (or dispersion) describes the spread or deviation of the individual values around the central value of a set of data. To illustrate this, let us consider the data given below :

Table 4.1.

Poultry A		*Poultry B*		*Poultry C*	
Daily Egg production		*Daily Egg production*		*Daily Egg production*	
4000		4050		3900	
4000	$\bar{X} = \frac{20000}{5}$	4025	$\bar{X} = \frac{20000}{5}$	2100	$\bar{X} = \frac{20000}{5}$
4000		3950		1200	
4000		3835		800	
4000	= 4000	4140	= 4000	12000	= 4000

Since, the mean egg production of A, B and C is the same, we are likely to conclude that the production pattern of the eggs is similar. But, in reality this is not true. In poultry A, daily egg production is the same irrespective of the day, whereas there is less amount of variation in the egg production for poultry B and greater amount of variation in the egg production for farm C. Therefore, different sets of data may have the same central value but differ greatly in terms of variation or dispersion.

Different measures of variation. There are following five different measures of variation :

1. Range
2. Quartile deviation (or semi-inter quartile range)
3. Mean deviation
4. Standard deviation

and 5. Variance.

1. RANGE

Range is defined as the difference between the highest value and lowest value in a set of data.

$$\boxed{\text{Range} = \text{Highest value} - \text{Lowest value}}$$

or $R = H - L$

Here R = Range
H = The highest value of the data
L = The lowest value of the data.

Computation of Range :

Ungrouped data. Range is obtained using above formula in ungrouped data.

Example. Hb% of 15 persons of a locality was observed as follows :
11.5, 11.7, 11.8, 12.5, 12.9, 13.8, 13.1, 14, 14.1, 14.2, 14.3, 14.5, 14.7, 14.8, 14.9 g/100 c.c. Find the range of the given data.

Sol. $R = H - L$

$\therefore$ $R = 14.9 - 11.5 = 3.5$. **Ans.**

Grouped data. For grouped data range is the difference between the upper true limit of the highest class and the lower true limit of the lowest class. Here H and L will be considered as upper true limit of highest class and lowest true limit of the lowest class respectively.

Example. Find the range of data given in following table :

Table 4.2.

No. of Pods	*True class limits*	*Frequency*
15–17	14.5–17.5	5
18–20	17.5–20.5	6
21–23	20.5–23.5	8
24–26	23.5–26.5	12
27–29	26.5–29.5	22
30–32	29.5–32.5	18
33–35	32.5–35.5	15
36–38	35.5–38.5	9
39–41	38.5–41.5	5

Sol. In this case, the upper true limit of the highest class 39–41 is 41.5 (H = 41.5) and the lower true limit of the lowest class 15–17 is 14.5 (L = 14.5)

$\because$ Range R = H – L

$\therefore$ R = 41.5 – 14.5 = 27. **Ans.**

The relative measure corresponding to range, called the co-efficient of range, is as follows :

$$\text{Co-efficient of range} = \frac{H-L}{H+L}$$

ere H = Highest value, L = Lowest value.

..erits and demerits of range :

Merits. It is easy to calculate and is useful in those cases whose variation of only two extreme points are considered. It is extensively used in statistical quality control. Range is helpful in studying the variations in the prices of shares and debentures and other commodities that are very sensitive to price changes from one period to another. It is also useful in meteorological weather forecast. In biological studies it can only indicate the variation in two extreme values of a data.

Demerits. It is crude measure of dispersion since it uses only two extreme values.

2. QUARTILE DEVIATION

Quartile deviation (or semi-interquartile range) is slightly better measure of dispersion than range. *"The range of variable between 25th percentile or 1/4th (First quartile–Q_1) and 75th percentile of 3/4th (third quartile–Q_3) divided by 2 is called quartile deviation or semi-interquartile range."* 25th percentile or 1/4 of a range of variable is denoted by Q_1 and 75th percentile or 3/4th of a range of variable is denoted by Q_3. When distance between Q_1 and Q_3 is divided by 2 then the obtained value of range is called Quartile deviation or Q.

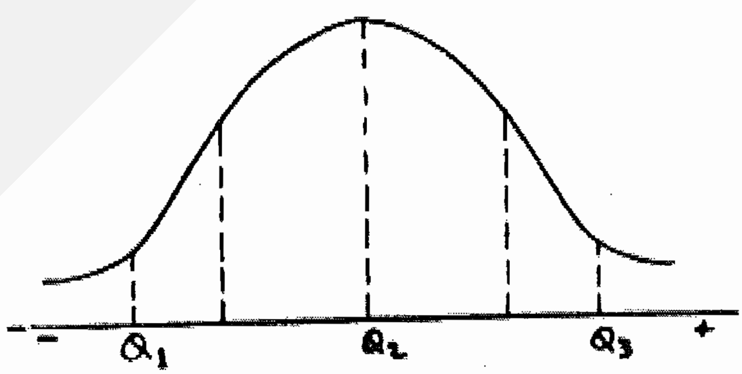


Fig. 4.1. Showing first Quartile (Q_1), Median (Q_2) and third Quartile Q_3).

We may understand quartile deviation with the help of Fig. 4.1.

The above figure shows that Q_1, Q_2 and Q_3 divide the distribution in four parts. Upto Q_1, is the first part (25%) of distribution. From Q_1 to Q_2, is the second part (25%) of distribution. Likewise from Q_2 to Q_3 and above Q_3 is the third and fourth part of distribution respectively. The distance between Q_1 and Q_3 is called interquartile deviation. Interquartile deviation divided by 2 is called semi-interquartile deviation. (In normal distribution it is called Probable error or PE).

Calculation of quartile deviation in ungrouped data. The following formula is used to calculate Q in ungrouped data.

$$Q = \frac{Q_3 - Q_1}{2} \quad \text{(ungrouped data)}$$

where
Q = Quartile deviation or semi-interquartile range
Q_1 = 1st quartile or 25th percentile
Q_3 = 3rd quartile or 75th percentile
Q_2 = Median.

Solved example. The range of a variable such as height is first quartile Q_1 (166.84 cm) and third quartile Q_3 (174.93 cm). Find the semi inter-quartile range or quartile deviation

According to formula :

$$Q = \frac{Q_3 - Q_1}{2}$$

$$= \frac{174.93 - 166.84}{2}$$

$$= \frac{8.09}{2} = 4.04 \text{ cm.}$$

Calculation of Q in grouped data : Formula to calculate quartile deviation *i.e.*, Q_1 and Q_3 (from grouped data) are as follows :

$$Q_1 = l + \frac{\left(\frac{N}{4} - F\right) \times i}{fq}$$

$$Q_3 = l + \frac{\left[\frac{3N}{4} - F\right] \times i}{fq}$$

Here l = Lower limit of that class interval where $Q_1\left[\frac{N}{4}\right]$ or $Q_3\left[\frac{3N}{4}\right]$ falls.

F = Cumulative frequency just above of that class interval where $Q_1\left[\frac{N}{4}\right]$ or $Q_3\left[\frac{3N}{4}\right]$ falls.

fq = Frequency of that class interval where Q_1 or Q_3 falls.

i = Length of class interval.

Solved example. Water percentage in the body of a species of fish and their frequency is given in following table. Compute the quartile deviation (Q).

Table 4.3.

Class-interval	*Frequency*	*Class-interval*	*Frequency*
16–20	4	41–45	3
21–25	3	46–50	3
26–30	8	51–55	2
31–35	9	56–60	2
36–40	14	61–65	2
			$\Sigma f = 50$

Calculation. First of all one should make a cumulative frequency table with the help of given data.

Table 4.4.

Class-interval	*Frequency*	*Cumulative frequency*
16–20	4	= 4
21–25	3	4 + 3 = 7
26–30	8	7 + 8 = 15, Q_1 lies within this class interval
31–35	9	15 + 9 = 24
36–40	14	24 + 14 = 38, Q_3 lies within this class interval
41–45	3	38 + 3 = 41
46–50	3	41 + 3 = 44
51–55	2	44 + 2 = 46
56–60	2	46 + 2 = 48
61–65	2	48 + 2 = 50

To find out Q we have to find the value of Q_1 and Q_3.

To find out Q_1 and Q_3 following formula have been used :

$$Q_1 = l + \frac{\left[\frac{N}{4} - F\right] \times i}{fq}$$

$$Q_3 = l + \frac{\left[\frac{3N}{4} - F\right] \times i}{fq}$$

Now before calculating Q_1 and Q_3 we have to find the value of $\frac{N}{4}$ and $\frac{3N}{4}$.

$$Q_1 = \text{Size } \frac{N}{4}\text{th observation} = \frac{50}{4} = 12.5.$$

The cumulative frequency in ascending order indicate that $\frac{N}{4}$th observation *i.e.*, 12.5 falls in class-interval 26–30. Therefore value of

$$Q_1 = 25.5 + \frac{(12.5-7)\times 5}{8}$$

$$= 25.5 + \frac{(5.5)\times 5}{8}$$

$$= 25.5 + (0.687)\times 5$$

$$= 25.5 + 3.437 = 28.93$$

$$Q_3 = \text{Size of } \frac{3N}{4}\text{th observation}$$

$$= \frac{3\times 50}{4} = 37.5$$

The cumulative frequency downward indicate that $\frac{3N}{4}$th observation or 37.5 falls in class-interval 36–40.

Now putting the value in formula :

$$Q_3 = \frac{(35.5-24)}{14}\times 5$$

$$= 35.5 + \left(\frac{13.5}{14}\right)\times 5$$

$$= 35.5 + (.964)\times 5$$
$$= 35.5 + (4.821) = 40.32.$$

After obtaining the values of Q_1 and Q_3, we can calculate Q with the help of following formula :

$$Q = \frac{Q_3 - Q_1}{2}$$

$$= \frac{40.32 - 28.93}{2}$$

$$= \frac{11.39}{2} = 5.69. \qquad \textbf{Ans.}$$

Co-efficient of quartile deviation :

$$\text{Co-efficient of Q} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{40.32 - 28.93}{40.32 + 28.93} = \frac{11.39}{69.25}$$

$$= 0.164. \quad \textbf{Ans.}$$

Merits and demerits of quartile deviation : The quartile deviation is superior to the range as it is not based on two extreme values but rather on middle 50% observations. Another advantage of quartile deviation is that it is the only measure of dispersion which can be used for open-end distribution.

3. MEAN DEVIATION

Synopsis. Mean deviation may be defined as "*the mean of all the deviations, in a given set of data obtained from an average.*" All the deviations are treated as positive.

[Every score or measurement deviates from the central value (Mean) and it is a certain distance above or below the mean value unless it happens to coincide with the mean, in which case the deviation is zero. Deviations above the mean are regarded as –ve dispersion, those below the mean as +ve dispersion.]

Mean deviation is calculated differently in ungrouped data and grouped data.

Ungrouped data. Following formula is used to obtain Mean deviation from ungrouped data :—

$$\text{M.D. or } \delta = \frac{\Sigma |x|}{N}$$

Here, M.D. or δ = Mean deviation;
x = Deviation from actual mean.
Σx = Sum of all deviations.
$||$ = Not considering signs (+ve or –ve) while summing up all deviations.

Deviation is obtained as follows :

$$x = \text{Score} - \text{Mean}$$
$$= X - \bar{X}$$

N = Number of observations.

(Here one thing is notable. Capital X is used for score or measurement of character and small x for deviation).

Example. Hb% of 10 patients of a ward of a hospital were obtained as 5, 7, 8, 10, 14, 12, 13, 5, 8, 8. Compute the Mean deviation.

Calculation. Following 4 steps have to be taken to calculate Mean deviation in ungrouped data—

(1) Find out the mean of the series.
(2) Find out the distance between each score and mean.
(3) Sum up all deviations. All deviations are treated as positive.
(4) Divide sum of all deviations by total number of observations.

Step 1. Mean $= \dfrac{\Sigma X}{N}$

$$= \frac{5+7+8+10+14+12+13+5+8+8}{10}$$

$$= \frac{90}{10} = 9.$$

Step 2. Following Table 4.5 is prepared to obtain deviations between each score and mean.

Table 4.5.

Score (X)	*Score—Mean* $(X - \overline{X})$	*Deviation* (x)
5	5–9	– 4
7	7–9	– 2
8	8–9	– 1
10	10–9	+ 1
14	14–9	+ 5
12	12–9	+ 3
13	13–9	+ 4
5	5–9	– 4
8	8–9	– 1
9	8–9	– 1

Step 3. Sum up all deviations regardless of sign.

$$\Sigma x = 4+2+1+1+5+3+4+4+1+1$$
$$= 26.$$

Step 4. $\text{MD} = \dfrac{\Sigma|x|}{N} = \dfrac{26}{10} = 2.6.$ **Ans.**

Grouped data. Following formula is used to obtain Mean deviation from grouped data :

$$\text{MD} \quad \text{or} \quad \delta = \frac{\Sigma|f.X|}{\Sigma f}.$$

Here $\Sigma f.\ X$ = Sum of multiplication of each frequency and each score.

Σf = Sum of all frequency.

Example. Testis weight (g) of 50 fishes of a species and their frequency were obtained as

$$\left[\frac{2}{2}; \frac{2.5}{1}; \frac{2.7}{1}; \frac{2.9}{2}; \frac{3}{3}; \frac{3}{1}; \frac{3.3}{3}; \frac{3.7}{2}; \frac{3.9}{4}; \frac{4}{3}; \frac{4.6}{2}; \frac{4.8}{3}; \frac{4.9}{3}; \frac{5}{3}; \frac{5.5}{2}; \frac{5.9}{3}; \frac{6}{3}; \frac{6.1}{3}; \frac{6.7}{3}; \frac{6.9}{3}\right]$$

Compute Mean deviation.

Calculation. Following Table 4.6 of 6 columns is prepared.

Column 1 of class interval. Length of class interval is kept 9.

Column 2 of Mid point of each class interval.

Column 3 of frequency of each class interval.

Column 4 of multiplication of mid point and frequency of each C.I.

Column 5 of deviation of each score from mean.

Column 6 Multiplication of deviation of each C.I. and frequency.

Table 4.6.

C.I.	*Mid Point X*	*Frequency f*	*Freq. MP f.X*	*Deviation x*	*Freq. deviation f.x*
2–2.9	2.45	6	14.7	−2.14	−12.84
3–3.9	3.45	13	44.85	−1.14	−14.82
4–4.9	4.45	11	48.95	−0.14	−1.54
5–5.9	5.45	8	43.6	+ 0.86	+ 6.88
6–6.9	6.45	12	77.4	+ 1.86	+ 22.32
	$\Sigma X = 22.25$	$\Sigma f = 50$	$\Sigma f.X = 229.5$		$\Sigma \mid f.X \mid = 58.4$

First of all obtain mean $X = \dfrac{\Sigma f.X}{\Sigma f} = \dfrac{229.5}{50} = 4.59.$

Then obtain deviation of each score from mean (mentioned in 5th column). It is obtained by $X - \overline{X}$.

Now frequency of each C.I. multiplied by each deviation and is finally sum of all is obtained.

$$\text{MD} \quad \text{or} \quad \delta = \frac{\Sigma \mid f.x \mid}{\Sigma f} = \frac{58.4}{50}$$

$$= 1.168. \quad \textbf{Ans.}$$

Coefficient of mean deviation may be calculated by following formula

$$\text{Coefficient of mean deviation} = \frac{\text{MD}}{\text{Mean}} = \frac{1.168}{4.59} = 0.254. \quad \textbf{Ans.}$$

Merits and Demerits of mean deviation :

Merits. It is easy to calculate.

Demerits. It is less reliable because positive and negative signs are ignored.

4. STANDARD DEVIATION

Standard deviation is the most important and widely used measure of dispersion. It is denoted by a Greek letter σ (sigma).

The standard deviation may be defined as "*the square root of the arithmetic mean of the squared deviations of measurements from their mean.*" It has accordingly often been called the *root mean-square deviation.*

Standard deviation is calculated differently in *ungrouped* and *grouped data.*

Ungrouped data :

Following formula is used where deviation is obtained from mean.

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \text{ or } \sqrt{\frac{\Sigma x^2}{N-1}}.$$

Here x = deviation obtained from actual mean.

N = Total number of observations.

Note : Standard deviation is computed by using N – 1 in the denominator of above formula in place of N if size of sample (*i.e.*, total number of observations) is less than 30. If size of sample is more than 30 then previous formula *i.e.*,

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$$

is used. For reasons consult Chapter 8 of the book, Fundamental Statistics in Psychology and Education by Guilford. 4th edition—McGraw-Hill Book Company, New York.

The above formula calls for following six steps in computation in fixed order :

Step 1. Find mean of the series.

Step 2. Find deviation of each score from the mean.

Step 3. Square each deviation, finding x^2.

Step 4. Sum the squared deviations, finding Σx^2.

Step 5. Divide this sum by N or N – 1, finding

$$\frac{\Sigma x^2}{N} \text{ or } \frac{\Sigma x^2}{N-1}.$$

Step 6. Extract the square root of the result of step 5. This is standard deviation.

Solved Example. Haemoglobin per cent g/100 ml. of *Heteropneustes fossils* was recorded as 23, 22, 20, 24, 16, 17, 18, 19 and 21. Compute the standard deviation (σ) by Indirect method.

Calculation. Following table 4.7 having four columns is prepared on the basis of above observations. As per *above steps* one has to find mean of the series. Here $\Sigma X = 180$ and number of observations

$$N = 9.$$

$$\therefore \quad \text{Mean} = \frac{180}{9} = 20.$$

Table 4.7.

Observation X	*Observation-Mean* $X - \overline{X}$	*Deviation* x	*(Deviation)²* x^2
16	16–20	– 4	16
17	17–20	– 3	9
18	18–20	– 2	4
19	19–20	– 1	1
20	20–20	0	0
21	21–20	+ 1	1
22	22–20	+ 2	4
23	23–20	+ 3	9
24	24–20	+ 4	16
$\Sigma X = 180$			$\Sigma x^2 = 60$

Here size of sample is less than 30. Therefore following formula is applicable.

$$\sigma = \sqrt{\frac{\Sigma x^2}{N-1}}$$

on putting the values in the above formula

$$\sigma = \sqrt{\frac{60}{9-1}}$$

$$= \sqrt{\frac{60}{8}} = \sqrt{7.5} = 2.75. \quad \textbf{Ans.}$$

Standard deviation from grouped data :

Following formula is used to obtain standard deviation by Long Method from grouped data :

$$\sigma = \sqrt{\frac{\Sigma f . x^2}{\Sigma f}}$$

The above formula calls for following steps in computation in fixed order.

Step 1. Find mid point of each class interval.

Step 2. Find mean value of the series using formula $\frac{\Sigma f \cdot X}{f}$.

Step 3. Find each deviation from the mean.

Step 4. Square each deviation, finding x^2.

Step 5. Multiply each squared deviation with corresponding frequency, finding $f. x^2$.

Step 6. Sum the squared deviation multiplied by frequency, finding $\Sigma f. x^2$.

Step 7. Divide the sum of step 6 *i.e.* $\Sigma f.x^2$ by Σf, finding $\frac{\Sigma f. x^2}{\Sigma f}$

Step 8. Extract the square root of the result of step 7.

Solved Example. Weight of testis of 50 frogs is given below with their frequency. Find the standard deviation.

Wt. of Testis	2	2.5	2.7	2.9	3	3.1	3.3	3.7	3.9	4
Frequency	2	1	1	2	3	1	3	2	4	3

Wt. of Testis	4.6	4.8	4.9	5	5.5	5.9	6	6.1	6.7	6.9
Frequency	2	3	3	3	2	3	3	3	3	3

Calculation. On the basis of above instruction Table 4.8 is prepared.

$$\text{Mean} = \frac{\Sigma f \cdot X}{\Sigma f} = \frac{229.5}{50} = 4.59.$$

Table 4.8. Here standard deviation is obtained with the help of Actual mean.

C.I.	*Mid point* X	*Frequency* f	$f.X$	*Deviation* x	*Deviation squared* x^2	$f.x^2$
2–2.9	2.45	6	14.7	– 2.14	4.5796	27.47
3–3.9	3.45	13	44.85	– 1.14	1.2996	16.88
4–4.9	4.45	11	48.95	– 0.14	0.0196	0.21
5–5.9	5.45	8	43.6	+ 0.86	0.7396	5.91
6–6.9	6.45	12	77.4	+ 1.86	3.4596	41.4
		$\Sigma f = 50$	$\Sigma f.X = 229.5$		$\Sigma x^2 = 10.088$	$\Sigma f. x^2 = 91.87$

Deviation x of each score (From mid point) obtained from this actual mean using formula $X - \overline{X}$. For instance deviation

$$x = X - \overline{X}$$
$$= 2.45 - 4.59 = - 2.14.$$

Compute the obtained value in following formula

$$\sigma = \sqrt{\frac{\Sigma f.x^2}{\Sigma f.}}$$

$$= \sqrt{\frac{91.87}{50}} = \sqrt{1.05} = 1.02 \quad \textbf{Ans.}$$

Merits and Demerits of Standard Deviation :

Merits :

(*i*) It summarises the deviation of a large distribution from mean in one figure used as a unit of variation.

(*ii*) It indicates whether the variation of difference of an individual from the mean is real or by chance.

(*iii*) It helps in calculating the standard error.

(*iv*) It helps in finding the suitable size of sample for valid conclusions.

Demerits : It gives weightage to only extreme values. The process of squaring deviations and then taking square root involves lengthy calculation.

5. VARIANCE

Introduction. V is used to stand for variance and it has algebraic interrelationship with standard deviation. The square of the standard deviation is called variance. This may be demonstrated symbolically as follows :

$$\sigma^2 = V.$$

Definitions. Variance may be defined as "*Square of sum of deviation divided by number of observations*" or *"The square of the standard deviation is termed as variance.*

Interpretations of variance. Suppose that we have a sample of only one case, with only one score. There is no possible basis for individual differences in such a sample, and therefore there is no variance or variability. Consider a second individual with his score in the same test or experiment. We now have one difference. Consider a third case and we then have two additional differences, three altogether. There are as many differences as there are possible pairs of individuals. We could compute all these interpair differences and could average them to get a

single, representative value. We could also square them and then average them. It is most easy to find a mean of all scores and to use that value as a common reference point.

Each difference then becomes a deviation from that reference point, and there are only as many deviations as there are individuals. Either the variance or the S.D. is a single representative value for all the individual differences when taken from a common reference point.

Solved example 1. Hb% of 10 patients of a ward was recorded as 7, 8, 9, 10, 11, 12, 13, 14.5, 15 and 15.5 g/100 ml. Find out the variance of the data.

Following table of three columns was prepared from above ungrouped data.

Table 4.9.

Hb% X	*Deviation* $X - \overline{X} = x$	*Standard deviation* x^2
7	$7 - 11.5 = -4.5$	20.25
8	$8 - 11.5 = -3.5$	12.25
9	$9 - 11.5 = -2.5$	6.25
10	$10 - 11.5 = -1.5$	2.25
11	$11 - 11.5 = -0.5$	0.25
12	$12 - 11.5 = +0.5$	0.25
13	$13 - 11.5 = +1.5$	2.25
14.5	$14.5 - 11.5 = +3.0$	9.0
15.0	$15 - 11.5 = +3.5$	12.0
15.5	$15.5 - 11.5 = +4.0$	16.0
$\Sigma X = 15.0$		$\Sigma x^2 = 80.75$

$$\overline{X} = \frac{\Sigma X}{N}$$

Here $\Sigma x = 115$, $N = 10$, $\overline{X} = \frac{115}{10} = 11.5$

$\therefore$ Variance or $V = \frac{\Sigma x^2}{N} = \frac{80.75}{10} = 8.075$. **Ans.**

Measurements of relative dispersion (coefficient of variatio
Measures of dispersion give us an idea about the extent to which variat are scattered around their central value. Therefore, two distributio having the same central values can be compared directly with the help various measures of dispersion. If, for example, an analysis of seed numb per fruit in two batches of 10 fruits in a garden, batch I have a *mean sco* $\overline{X}_1 = 70$ with S.D (σ_1) = 1.25 and batch II have a *mean score* $\overline{X}_2 = 80$ wi SD (σ_2) = 2.4 then it is clear that batch I having a lesser value of SD (are more consistent in producing seed number than batch II.

On the other hand we have situations when two or more distributions having unequal means or different units of measurements are to be compared in respect of their variability. For making such comparisons we use a statistic called *relative dispersion* or coefficient of variation (*c.v.*). Formula of coefficient of variation is as follows :

$$c.v. = \frac{\sigma}{\overline{X}} \times 100 \quad \text{or} \quad c.v. = \frac{100\,\sigma}{\overline{X}}$$

Solved example 1. An analysis of seed number per fruit in 10 fruits each of two batches is given below (Table 4.10). Find c.v. of both batches and mention which of the two has lower range of variation.

Table 4.10.

Fruit No.	*No. of seeds*	
	Batch I (X_1)	*Batch II* (X_2)
1	7	10
2	9	8
3	6	9
4	8	10
5	6	11
6	5	10
7	7	5
8	8	6
9	6	4
10	8	7
	$\Sigma X_1 = 70$	$\Sigma X_2 = 80.$

Here both $N_1 = 10$ and $N_2 = 10$.

$\Sigma X_1 = 70$ and $\Sigma X_2 = 80$.

(Calculate Mean and S.D. of both batches)

$\overline{X}_1 = 7$ and $\overline{X}_2 = 8$

$\sigma_1 = 1.25$ and $\sigma_2 = 2.4$

$$\because \quad cv = \frac{100\,\sigma}{\overline{X}}$$

$$\therefore \quad cv_1 = \frac{100 \times 1.25}{7} = 17.8\%$$

$$cv_2 = \frac{100 \times 2.4}{8} = 30\%.$$

Deduction. Fruits of Batch I are more consistent in seed production than Batch II.

Solved example 2. Mean values of Hb% of 20 males and 20 females were calculated as 13.5 and 14 mg/100 ml. SD of males as 3 and 4 respectively. Find coefficient of variation of both male and female. Mention which sex is more variable and which more consistent.

Table 4.11.

Group of sex	*Mean*	*S.D*
Males	13.5	3
Females	14	4

For males $\quad cv = \dfrac{100 \times 3}{13.5} = 22.22\%.$

For females $\quad cv = \dfrac{100 \times 4}{14} = 28.57\%$

Deduction. Females are variable than males in respect of Hb%. In other words contrary to females, males are more consistent in Hb%.

Solved example 3. Compare the relative variability of the following parameters of a species of fish from the following data :

(1) Mean of total length $\overline{X}_1 = 15$, S.D. = 1.5

(2) Mean standard length $\overline{X}_2 = 12.5$, S.D. = 0.5.

(3) Mean of length from Snout to Pectoral fin $\overline{X}_3 = 5.0$ S.D. = 0.3

Calculations :

(1) cv of total length $= \dfrac{100 \times 1.5}{15} = 10\%.$

(2) cv of standard length $= \dfrac{100 \times 0.5}{12.5} = 4\%.$

(3) cv of Snout to Pectoral fin $= \dfrac{100 \times 0.3}{5} = 6\%.$

Conclusion. (1) Total length is the most variable parameter as the coefficient of variation is highest for this *i.e.*, 10%.

(2) Standard length is the least variable parameter as the cv is lowest *i.e.*, only 4%.

(3) Length from Snout to Pectoral fin is medium as the cv is in between first and second *i.e.*, 6%.

Merits and demerits of variance :

(1) It is easy to calculate (2) It indicates the variability clearly. But the use of cv creates two difficulties —

(*a*) The unit of expression of variance is not the same as that of the observations, because variance indicates squared deviations. For example if '*x*' values are obtained in cm variance will be in square cm.

(*b*) Variance is usually a large number of compared to the values of observations. Therefore variance is now seldom used to express the variability.

EXERCISE

1. What do you mean by dispersion or deviation or scatter or variability ? How many types of variability are there ? Describe them in brief.
2. What do you mean by measures of variability ? Name measures of variability of individual observations.
3. Define Range. Find out Range of following series :
   (*i*) 60, 72, 81, 5, 70, 72, 78, 66, 55, 58, 90.
   (*ii*) 11, 11.5, 18.9, 17.2, 14, 18, 16, 16.2, 16.2, 13.2, 22.4.
4. Quartile deviations ? Mention the formula of Quartile deviation (Q) and calculate it with the help of following data :

Class-interval	*Frequency*
30–39	6
40–49	8
50–59	9
60–69	12
70–79	7
80–89	4
90–99	2

5. (*i*) What do you mean by deviation ? Calculate mean deviation of following data both in ungrouped and grouped condition.
   Water % in the body of 15 fishes of a species is
   62, 62, 63, 64, 67, 68, 71, 72, 71, 69, 68, 64, 62, 66, 68. **[Ans.** 2.97]
   (*ii*) 60, 72, 81, 5, 70, 72, 78, 66, 55, 58, 90. **[Ans.** 13.60]
   (*iii*) 11, 11.5, 18.9, 17.2, 14, 18, 16, 16.2, 13.2, 22.4. **[Ans.** 2.77]
6. Find mean deviation from the given data in Q. 4. **[Ans.** 13.52]
7. Standard deviation ? Why standard deviation is popular than mean deviation in Biological statistical analysis ?
8. Calculate standard deviation and variance in the following set of data :

No. of flowers/plant (*X*)	*Frequency* (*f*)
1–3	6
4–6	14
7–9	9

9. Calculate mean deviation, Standard deviation and variance for the data relating to pH of the water sample.

pH water sample

6.6, 6.8, 6.1, 7.2, 7.1, 7.3, 7.4, 7.5.

10. Calculate the coefficient of variance in the following two sets of data. Show which one of the two has higher rate of variation.

Set I No. of cell/ml $(10)^3$	*Set II* No. of cells/ml $(10)^3$
6	8
5	9
8	10
6	3
4	2
4	9
5	6
3	3
4	4

Q. 11. Calculate mean deviation, standard deviation and variance in the two sets of data (control and treated). The data relate to increase in dry weight expected to a particular treatment dose and a control.

Observation No.	*Increase in dry weight (mg)*	
	Control Set Increase in dry weight (mg)	*Treated Set Increase in dry weight (mg)*
1	2.85	4.25
2	2.90	4.20
3	2.75	4.15
4	3.05	3.35
5	3.30	3.25
6	2.90	4.70
7	2.95	3.25
8	3.50	3.75
9	3.45	3.70
10	2.95	3.90

5. Tests of Significance

Synopsis. *Introduction, standard error of mean, standard error of standard deviation, student's t-test, chi-square test.*

Introduction. By significance of statistics we mean non-chance difference between obtained scores on the basis of sample and scores based on some hypothesis. If observed difference is significant then we say that observed difference is not influenced by chance defying null hypothesis. On the other hand if observed difference is not significant then one can say that observed difference is obtained by chance. Here we shall deal only one method of test of significance known as "*standard error*" and will learn about the use of students' *t*-test and chi-square (χ^2) test.

Standard error. Standard error is a measure or chance variation and it is not an error or mistake. Theoretically the difference between mean of population and mean of sample should be zero. But in biological experiments standard error is never zero. The statement may be understood with the help of following example.

Suppose in a poultry farm total number (Population) of hen is 1500. Every day average egg, laying capacity of all hen is 105. 200 hen (sample) were selected randomly from the population. Per day the average number of egg lying by these samples (200 hens) comes to 100. Here difference between observed mean (105) and expectation (100) comes to (105–100) = 5. This non-chance difference in between two means is called standard error.

Uses of standard error of mean

(*i*) To work out the limits within which the population mean would lie.

(*ii*) To determine whether the sample is drawn from a known population or not, when its mean is known.

(*iii*) To determine the standard error of difference between two means to know if the observed difference between the means of two samples is real and statistically insignificant or it is apparent and insignificant due to chance.

(*iv*) To calculate the size of sample.

Here we shall deal with

1. **Standard error of mean and**
2. **Standard error of standard deviation.**

1. **Standard error of mean.** Standard error of mean is the ratio of standard deviation of the sample divided by the square root of the total number of observations.

$$SE_M = \frac{\text{Standard deviation}}{\sqrt{N}} \quad \text{or} \quad \frac{\sigma}{\sqrt{N}}$$

Here, SE_M = Standard error of mean,
σ = Standard deviation
$\sqrt{N}$ = Square root of the total number of observations.

SE_M in ungrouped data. Following steps have to be taken to obtain SE_M in ungrouped data :

— Find standard deviation (σ) using following formula

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \quad \text{or} \quad \sqrt{\frac{\Sigma x^2}{N-1}}.$$

— Compute the value of Standard deviation and N in following formula :

$$SE_M = \frac{\sigma}{\sqrt{N}}.$$

Example. Size of 5 fishes in cm are 2, 5, 3, 4, 1, respectively. Find standard error of mean.

Calculation. Find standard deviation using following formula

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \quad \text{or} \quad \sqrt{\frac{\Sigma x^2}{N-1}}. \qquad \text{(ungrouped data)}$$

To obtain standard deviation following steps have to be taken :

— Find the mean of the series formula

$$\overline{X} = \frac{\Sigma X}{N}$$

Here $$\overline{X} = \frac{2+5+3+4+1}{5} = \frac{15}{5} = 3.$$

— Find deviations of the individual measurements from the mean.

Table 5.1.

Size of fish *X*	$X - \overline{X}$	*Deviation* *x*	x^2
2	2—3	−1	1
5	5—3	+2	4
3	3—3	0	0
4	4—3	+1	1
1	1—3	−2	4
			$\Sigma x^2 = 10$

— Find the sum of squares of deviations of individual measurements from their mean. In above example $\Sigma x^2 = 10$.

— Find standard deviation (σ) using following formula :

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \quad \text{or} \quad \sqrt{\frac{\Sigma x^2}{N-1}} \quad \left(\text{Here} \quad \sqrt{\frac{\Sigma x^2}{N-1}}\right.$$

formula is used because no. of observation is less than 30).

$$\sigma = \sqrt{\frac{10}{5-1}} = \sqrt{\frac{10}{4}}$$

$$= \sqrt{2.5} = 1.5.$$

Put the value of standard deviation in the formula

$$SE_M = \frac{\sigma}{\sqrt{N}} \quad \text{(ungrouped data)}$$

$$SE_M = \frac{1.5}{\sqrt{5}} = \frac{1.5}{2.23} = 0.67.$$

Normally we express reliability of statistic in the form of confidence interval. For example .95 and .99 are two very popular confidence level.

Taking into consideration above solved example, we may say that to know the divergence of the sample mean from population mean we multiply SE_M with that of ± 1.96 ($\pm 1.96 \times SE_M$).

Here, $SE_M = .67$

One can draw inference that obtain mean (3) is repeated 95 times out of 100 times ($\pm 1.96 \times .67$) = ± 1.31 upto unit (*i.e.* 1.96 to 4.35) may differ from population mean which is negligible. Therefore we can say that obtained mean is reliable.

SE_M in grouped data. Following steps have to be taken to obtain SE_M in grouped data :

— Find standard deviation using following formula :

$$\sigma = \sqrt{\frac{\Sigma f \cdot x^2}{\Sigma f}} \quad \text{or} \quad \sqrt{\frac{\Sigma f \cdot x^2}{\Sigma f - 1}} \quad \text{(grouped data)}$$

— Compute the value of σ and f in the following formula :

$$SE_M = \frac{\sigma}{\sqrt{\Sigma f}} \quad \text{(grouped data)}$$

Example. Frequency of weight of 50 fishes of same species ranging 30—129 gm is given below. Calculate the standard error of the given data (grouped series).

Wt. in gm	*Frequency*	*Wt. in gm*	*Frequency*
30–39	5	80–89	10
40–49	4	90–99	4
50–59	3	100–109	3
60–69	8	110–119	2
70–79	9	120–129	2

Calculation. For standard error of mean (SE_M) we have to obtain standard deviation (σ). Standard deviation (σ) is calculated with the help of following formula :

$$\sigma = \sqrt{\frac{\Sigma f \cdot x^2}{\Sigma f - 1}} \quad \text{(grouped data)}$$

A table 5.2 of seven columns is made —

Table 5.2.

Variable C.I.	*Mid-point X*	*Frequency (f)*	*f.X*	$X - \overline{X} = x$	x^2	$f.x^2$
30–39	34.5	5	172.5	34.5–74.17 = –39.67	1573.70	51573.7
40–49	44.5	4	178	44.5–74.17 = –29.67	880.30	3521.2
50–59	54.5	3	163	54.5–74.17 = 19.67	386.90	1160.7
60–69	64.5	8	516	64.5–74.17 = 9.67	93.50	748.0
70–79	74.5	9	670.5	74.5–74.17 = 0.33	0.1089	1.701
80–89	84.5	10	845	84.5–74.17 = 10.33	106.70	1067.0
90–99	94.5	4	378	94.5–74.17 = 20.33	413.30	1653.2
100–109	104.5	3	313.5	104.5–74.17 = 30.33	919.90	2759.7
110–119	114.5	2	223	114.5–74.17 = 37.33	1393.52	2787.04
120–129	124.5	2	249	124.5–74.17 = 50.55	2533.10	5056.2

Here mean or $\overline{X} = \dfrac{\Sigma f \cdot X}{\Sigma f}$ (grouped data)

$$= \frac{3708.5}{50} = 74.17.$$

deviations of each individual observation is obtained in 5th column. $(X - \overline{X})$ and other sum such as $\Sigma x \cdot x^2$ is obtained.

$$\Sigma = \sqrt{\frac{\Sigma f \cdot x^2}{\Sigma f - 1}}$$

$$= \sqrt{\frac{70338.44}{50 - 1}}$$

$$= \sqrt{\frac{70338.44}{49}}$$

$$= \sqrt{1435.47} = 37.88.$$

Put the value of σ and $\sqrt{f}$ in following formula—

$$SE_M = \frac{\sigma}{\sqrt{\Sigma f}} \quad \text{(grouped data)}$$

$$= \frac{37.88}{\sqrt{50}}$$

$$= \frac{37.88}{7.07} = 5.35. \quad \textbf{Ans.}$$

One can draw inference that obtained mean (74.14) is repeated 95 times out of 100 times ($\pm 1.96 \times 5.36$) = ± 10.48 upto unit (67.69 to 84.65) may differ from population mean which is negligible. Therefore we can say that obtained mean is reliable.

2. **Standard error of Standard deviation.** Like mean, standard deviation (σ) also varies if obtained from different samples of same population. Therefore we can test the reliability of standard deviation by standard error.

Standard error of standard deviation is calculated differently in ungrouped and grouped data.

SEσ in ungrouped data

Following formula is used to obtain standard error of standard deviation in ungrouped data :

$$SE\sigma = \frac{\sigma}{\sqrt{2N}} \quad \text{(ungrouped data)}$$

Here SEσ = Standard error of standard deviation
σ = Standard deviation
N = Number of observations.

Example. Size of 5 fishes in cm are 2, 5, 3, 4, 1 respectively obtained standard error of standard deviation of the given ungrouped data. (σ is calculated in previous page of same data).

Put the value in following formula :

$$SE\sigma = \frac{\sigma}{\sqrt{2N}} \quad \text{(ungrouped data)}$$

$$= \frac{1.5}{\sqrt{2 \times 5}} = \frac{1.5}{\sqrt{10}}$$

$$= \frac{1.5}{1.58} = 0.95$$

Significance. Standard error = 0.95. It means σ(1.5) is repeated 95 times out of 100 times at 0.05 level. Therefore ($\pm 1.96 \times 0.95$) = ± 1.86

unit may deviate from standard deviation of population which very close to obtained standard deviation (1.5). Therefore we can say that obtained standard deviation is reliable, because there is not much variability in it.

SEσ in grouped data

Following formula is used to obtained SEσ in grouped data :

$$SE\sigma \frac{\sigma}{\sqrt{2.f}} \quad \text{(grouped data)}$$

Example. Testis weight of 50 fishes of a species of fish and their frequency were measured as follows. Find SEσ of the given grouped data :

$$\frac{\text{Weight of testis}}{\text{frequency}} \quad \frac{2}{2}, \frac{2.5}{1}, \frac{2.7}{1}, \frac{2.9}{2}, \frac{3}{3}, \frac{3.1}{1}, \frac{3.3}{3}, \frac{3.7}{2}, \frac{3.9}{4}, \frac{4}{3}, \frac{4.6}{2}, \frac{4.8}{3},$$

$$\frac{4.9}{3}, \frac{5}{3}, \frac{5.5}{2}, \frac{5.9}{3}, \frac{6}{3}, \frac{6.1}{3}, \frac{6.7}{3} \text{ and } \frac{6.9}{3}.$$

Sol. A table 5.3 of seven columns is prepared on the basis of above data to obtain SEσ using actual mean.

Table 5.3.

Class-interval	*Frequency* f	*Mid-point* X	*Freq. XMP* $f.X$	$X - \overline{X}$ x	x^2	$f.x^2$
2–2.9	6	2.45	14.7	−2.14	4.5796	27.47
3–3.9	13	3.45	44.85	−1.14	1.2996	16.89
4–4.9	11	4.45	48.95	−0.14	0.0196	0.21
5–5.9	8	5.45	43.6	+0.86	0.7396	5.91
6–6.9	12	6.45	77.4	+1.86	3.4596	41.51

Find actual mean of data

$$\overline{X} = \frac{\Sigma f . x}{\Sigma f} = \frac{229.5}{50} = 4.59.$$

Find deviation of each individual observation.

Multiply each deviation with frequency and sum them up. It comes to $\Sigma f . x^2 = 92.02$.

Find standard deviation with the help of following formula :

$$\Sigma = \sqrt{\frac{\Sigma f . x^2}{\Sigma f}} \quad \text{(grouped data)}$$

$$= \sqrt{\frac{92.02}{50}} = \sqrt{1.84} = 1.35$$

Obtain SEσ by putting the value of σ and frequency in following formula :

$$SE\sigma = \frac{\sigma}{\sqrt{2 . f}}$$

$$= \frac{1.35}{\sqrt{2.50}} = \frac{1.35}{\sqrt{100}}$$

$$= \frac{1.35}{10} = 0.135. \quad \textbf{Ans.}$$

Significance. Standard error of standard deviation = .14.

It means standard deviation (1.35) is repeated 95 times out of 100 (at 0.05 level) times ($\pm 1.96 \times 0.14$) = ± 0.27 unit may deviate from standard deviation of population. Therefore one can say that obtained standard deviation is reliable because there is not much variability in it.

Standard error of standard deviation ($SE\sigma$) may be obtained from grouped data using assumed mean. Following formula is used to obtain standard deviation applying assumed mean

$$\sigma = i \times \sqrt{\frac{\Sigma f . x'^2 - c^2}{\Sigma f}}$$

Example. Testis weight of 5 fishes of a species of fish and their frequency were measured as follows :

$$\frac{\text{Weight of testis}}{\text{Frequency}} \quad \frac{2}{2}, \frac{2.5}{1}, \frac{2.7}{1}, \frac{2.9}{2}, \frac{3}{3}, \frac{3.1}{1}, \frac{3.3}{3}, \frac{3.7}{2}, \frac{3.9}{4}, \frac{4}{3}, \frac{4.6}{2}, \frac{4.8}{3},$$

$$\frac{4.9}{3}, \frac{5}{3}, \frac{5.5}{2}, \frac{5.9}{3}, \frac{6}{3}, \frac{6.1}{3}, \frac{6.7}{3} \text{ and } \frac{6.9}{3}.$$

Calculation. A table of six columns is prepared to obtain standard error of standard deviation applying assumed mean.

Table 5.4.

C.I.	*Frequency* f	*M.P.* X	*Assumed mean* x'	$f.x'$	$f.x'^2$
2–2.9	6	2.45	–2	–12	24
3–3.9	13	3.45	–1	–13	13
				–25	
4–4.9	11	4.45	0	0	0
5–5.9	8	5.45	+1	+8	8
6–69	12	6.45	+2	+24	48
				+32	
				$\Sigma f . x^2 = 7$	$\Sigma f . x'^2 = 93$

$$c = \frac{\Sigma f \,.\, x'}{\Sigma f} = \frac{7}{50} = 0.14$$

$$c^2 = (0.14)^2$$
$$= 0.0196$$
$$i = 1$$

$$\because \quad \sigma = i \times \sqrt{\frac{\Sigma f \,.\, x'^2}{\Sigma f} - c^2}$$

$$= 1 \times \sqrt{\frac{92}{50} - 0.0196}$$

$$= 1 \times \sqrt{1.86 - 0.0196} = 1 \times \sqrt{1.8404}$$

$$= \sqrt{1.8404} = 1.356$$

Put the value of σ and f in following formula :

$$SE\sigma = \frac{\sigma}{\sqrt{2 \,.\, f}} = \frac{1.356}{\sqrt{2 \times 50}}$$

$$= \frac{1.356}{\sqrt{100}} = \frac{1.356}{10} = 0.1356. \quad \textbf{Ans.}$$

Standard error of standard deviation comes to about 0.1356. It means standard deviation (1.356) is repeated 95 times out of 100 (at .05 level) times $\pm 1.96 \times 0.1356 = 0.2657$ unit can deviate from standard deviation of population. This SEσ is reliable because it has got less variability.

EXERCISE

1. What do you mean by test of significance of a mean or difference between two means ? Define standard error. Explain and mention formula to obtain standard error of mean and standard error of standard deviation both in ungrouped and grouped data.
2. Rate of oxygen consumption in 15 groups of fishes of macrograthus was studied as 62, 64, 67, 68, 61, 72, 71, 69, 67, 64, 62, 60, 62, 64 and 67 kg/hour/100c.c. Obtain standard error of mean and standard error of standard deviation both in ungrouped and grouped series.

6. Student's t-Tests

Synopsis. *Introduction, Definition and formula, Significance, Application of t-test, unpaired t-test, Paired t-test, t-test for grouped data.*

Introduction. W.S. Gossett (1908) applied a statistical tool called *t*-test, to test the significance of the difference between two means. The pen name of Mr. Gossett was **Student** and therefore *t*-test is known as 'student's *t*-test'. Later on R.A. Fisher developed *t*-test and explained it in various ways.

Student's *t*-test is used not only to test the significance of difference between two means but also to test the significance of product moment correlation, point-biserial correlation, rank difference correlations etc.

Student's *t*-test is also known as *t*-ratio because it is ratio of difference between two means and standard error of difference between two means. Following formula is used to obtain *t*-ratio.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{SE_D} \quad \text{or} \quad \frac{\text{Diff. of two means}}{\text{SE of diff. between two means}}$$

Here, $\overline{X}_1$ = Mean of one variable

$\overline{X}_2$ = Mean of second variable

SE_D = Standard error of difference between two means.

To determine the significance. Probability of occurrence of any calculated value of '*t*' is determined by comparing it with the value given in the '*t*' table corresponding to the degree of freedom derived from the number of conservations in the sample under study. If the calculated value of '*t*' exceeds the value given in the *t* table (appendix 3) at different levels (.01, .05 etc.), it is said to be significant. But if calculated value of '*t*' is less than the table value then the difference between two means is insignificant.

Application of *t*-test. The significance of difference between two means is obtained differently in uncorrelated or unpaired *t*-test and paired *t*-test.

Unpaired or uncorrelated *t*-test. Unpaired *t*-test is applied to unpaired data of independent observations made on individuals of two different or separate groups or sample drawn from two populations.

[**Note :** Standard error of the difference between two uncorrelated means is calculated differently during *t*-test.]

Following steps have to be taken to test the significance of difference between two uncorrelated means :

— Find the observed difference between means of two samples $(\overline{X}_1 - \overline{X}_2)$.

— The standard error of the difference between uncorrelated sample means is obtained with the help of following formula :

$$SE_D = \sqrt{SE\overline{X}_1^2 + SE\overline{X}_2^2}$$

Here SE_D = Standard error of the difference between the two sample means.

$SE\overline{X}_1$ = Standard error of the first mean.

$SE\overline{X}_2$ = Standard error of the second mean.

We find SE_M of each mean with the help of following formula

$$SE\sigma = \frac{\sigma}{\sqrt{N}} \quad \text{or} \quad \frac{\sigma}{\sqrt{N-1}}.$$

To obtain $SE\sigma$ one have to obtain the value of combined (σ). Following formula is used to obtain σ.

$$\text{combined } \sigma = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}}.$$

Example. Body length of 10 fishes of a species of fish was obtained from two ponds (population) of Gaya town. They were measured as :

Pond A. 20 cm, 24 cm, 20 cm, 28 cm, 22 cm, 20 cm, 24 cm, 32 cm, 24 cm and 26 cm.

Pond B. 12 cm, 10 cm, 8 cm, 10 cm, 6 cm, 4 cm, 14 cm, 20 cm, 10 cm and 6 cm.

Calculate the mean difference in total body length between the two ponds of fish is significant or not.

Calculation. Following steps have to be taken to obtain *t* ratio :

— Make a table of six columns.

— First column having length of sample A ; Second column for $X_1 - \overline{X}_1 = x_1$; Third column $(X_1 - \overline{X}_1)^2$ or x_1^2. Fourth column having length of sample B ; Fifth column for $(X_2 - \overline{X}_2) = x_2$; Mean value ; Sixth column for $(X_2 - \overline{X}_2)^2$ or x_2^2.

Table 6.1.

Length of fish pond A X_1	*Deviation* $X_1 - \overline{X}_1 = x_1$	*Deviation*^{2} x_1^2	*Length of pond B* X_2	*Length-mean* $X_2 - \overline{X}_2 = x_2$	*(Length-Mean)*^{2} x_2^2
20	−4	16	12	2	4
24	0	0	10	0	0
20	−4	16	8	−2	4
28	4	16	10	0	0
22	−2	4	6	−4	16
20	−4	16	4	−6	36
24	0	0	14	4	16
32	8	64	20	10	100
24	0	0	10	0	0
26	2	4	6	4	16
$\Sigma X_1 = 240$		$\Sigma x_1^2 = 136$	$\Sigma X_2 = 100$		$\Sigma x_2^2 = 192$

Here, $$\overline{X}_1 = \frac{\Sigma X_1}{N_1} = \frac{240}{10} = 24$$

$$\overline{X}_2 = \frac{\Sigma X_2}{N_2} = \frac{100}{10} = 10$$

$$N_1 = 10 \quad \text{and} \quad N_2 = 10$$

$$\text{Pooled or combined } \sigma = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}}$$

$$= \sqrt{\frac{136 + 192}{10 + 10 - 2}}$$

$$= \sqrt{\frac{328}{18}} = 4.238.$$

Difference of two means

$$= \overline{X}_1 - \overline{X}_2 = 24 - 10 = 14$$

$$SE_D = \text{pooled } \sigma \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

and $$t = \frac{\overline{X}_1 - \overline{X}_2}{SE_D}$$

$$= \frac{14}{4.268\sqrt{\frac{1}{10}+\frac{1}{10}}}$$

$$= \frac{14}{4.268 \times 0.447} = \frac{14}{1.908}$$

$$= 7.337.$$

Significance. Here d.f. = 10 + 10 − 2 = 18.

Calculated value of t comes of 7.337. Table value of t at d.f. 18 on .01 level is 2.552. The calculated value is much greater than the tabulated value. Therefore the difference between two means of two populations (length of fishes of Pond A and B) is highly significant.

Example. 13 boys were given usual diet plus honey—while the second comparable group of 12 boys were given usual diet only. After one year the gain in weight in kg (both group) was noted. The noted weight in both groups is given below :

Can we say that honey was responsible for any difference in weight. Justify your answer using student's t–test.

Group X_1 of 13 boys. Increase in weight kept on usual diet plus honey : 5, 3, 4, 3, 2, 6, 3, 2, 3, 6, 7, 5 and 3 kg.

Group X_2 of 12 boys. Increase in weight kept on usual diet : 1, 3, 2, 4, 2, 1, 3, 4, 3, 2, 2, 3 kg.

Calculation. Following Table 6.2 of six columns is prepared to obtain t value.

Table 6.2.

X_1	$X_1 - \bar{X}_1 = x_1$	x_1^2	X_2	$X_2 - \bar{X}_2 = x_2$	x_2^2
5	1	1	1	−1.5	2.25
3	−1	1	3	0.5	0.25
4	0	0	2	−0.5	0.25
3	−1	1	4	1.5	2.25
2	−2	4	2	0.5	0.25
6	2	4	1	1.5	2.25
3	−1	1	3	.5	0.25
2	−2	4	4	1.5	2.25
3	−1	1	3	0.5	0.25
6	2	4	2	0.5	0.25
7	3	9	2	0.5	0.25
5	1	1	3	0.5	0.25
3	1	1			
$\Sigma X_1 = 52$		$\Sigma x_1^2 = 32$	$\Sigma X_2 = 30$		$\Sigma x_2^2 = 11.00$

Here, $\overline{X}_1 = \frac{52}{13} = 4$ and $\overline{X}_2 = \frac{30}{12} = 2.5$

For group X_1, $N_1 = 13$

and for group X_2, $N_2 = 12$

$\Sigma(X_1 - \overline{X}_1)$ or $x_1^2 = 32$

and $\Sigma(X_2 - \overline{X}_2)$ or $x_2^2 = 11$

$$\text{pooled } \sigma = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}}$$

$$= \sqrt{\frac{32 + 11}{13 + 12 - 2}}$$

$$= \sqrt{\frac{43}{23}} = \sqrt{1.869} = 1.367$$

Difference between two means

$$= \overline{X}_1 - \overline{X}_2 = 4 - 2.5$$

$$= 1.5$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{SE_D}$$

$$\therefore \quad SE_D = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

$$\therefore \quad t = \frac{1.5}{1.367 \times \sqrt{\frac{1}{13} + \frac{1}{12}}}$$

$$= \frac{1.5}{1.367} \times \sqrt{0.16}$$

$$= \frac{1.5}{1.367 \times 0.4} = 2.739$$

Here $d.f. = N_1 = N_1 + N_2 - 2$
$= 13 + 12 - 2 = 23.$

Significance. The calculated value of t comes to 2.739. The table value of t at d.f. 23 on .01 level is 2.500. The calculated value is bit higher, therefore we can say that honey has definite influence on the growth of children in positive direction.

Example. The ESR (erythrocyte) sedimentation rate (mm/hour) of 15 male and 10 female is given below. Calculate the significance of the difference in the means.

Males : 65, 60, 115, 82, 43, 103, 125, 118, 83, 75, 90, 95, 128, 65 and 84.

Females : 63, 85, 90, 100, 90, 105, 98, 93, 100, 125.

Calculation. Here, $N_1 = 15$ and $N_2 = 10$.

$\Sigma X_1 = 1331$ and $\Sigma X_2 = 949$.

$$\overline{X}_1 = \frac{1331}{15} = 88.73 \quad \text{and} \quad \overline{X}_2 = \frac{949}{10} = 94.9$$

Now to obtain standard deviation (σ_1 and σ_2) following table of 6 columns is prepared.

Table 6.3.

X_1	$X_1 - \overline{X}_1 = x_1$	x_1^2	X_2	$X_2 - \overline{X}_2 = x_2$	x_2^2
65	−23.73	563.11	63	−31.9	1017.61
60	−28.73	825.41	85	−9.9	98.01
115	+26.27	690.11	90	−4.9	24.01
82	−6.73	45.29	100	5.1	26.01
43	−45.73	2091.23	90	−4.9	24.01
103	14.27	203.63	105	10.1	102.01
125	36.27	1315.51	98	3.1	9.61
118	29.27	856.73	93	−1.9	3.61
83	−5.73	32.83	100	5.1	26.01
75	−13.73	188.51	125	30.1	906.01
90	1.27	1.61			
95	6.27	39.31			
128	39.27	1542.13			
65	23.73	563.11			
84	4.73	22.37			
		$\Sigma x_1^2 = 8980.86$			$\Sigma x_2^2 = 2236.9$

$$\sigma_1 \text{ or } SD_1 = \sqrt{\frac{\Sigma x^2}{N_1}} = \sqrt{\frac{8980.86}{15}}$$

$$= \sqrt{598.72} = 24.46$$

$$\sigma_2 \text{ or } SD_2 = \sqrt{\frac{\Sigma x^2}{N_2}} = \sqrt{\frac{2236.9}{10}}$$

$$= \sqrt{223.69} = 14.95.$$

$$\therefore \quad \text{Combined SD} = \sqrt{\frac{\Sigma(X_1 - \overline{X}_1)^2 + \Sigma(X_2 - \overline{X}_2)^2}{N_1 + N_2 - 2}} \quad \text{or} \quad \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2 - 2}}$$

$$= \sqrt{\frac{8980.86 + 2236.9}{25 - 2}}$$

$$= \sqrt{\frac{11217.76}{23}}$$

$$= \sqrt{487.73} = 22.08$$

$$SE_D = \sigma \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

$$= 22.08 \sqrt{\frac{1}{15} + \frac{1}{10}}$$

$$= 22.08 \times 0.408 = 9.01$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{SE\,(\overline{X}_1 - \overline{X}_2)}$$

$$= \frac{88.73 - 94.9}{9.01} = 0.68.$$

Conclusion : On perusal of 't' table at 15 + 10 – 2 = 23 d.f., the probability of occurrence of calculated value by chance is greater than 0.4. This leads to an acceptance to null hypothesis. Hence the observed difference between the mean ESR_S of male and female filarial patients is statistically insignificant.

Paired or correlated *t*-test. (The significance of the difference between two correlated means). Paired *t*-test is applied to paired data obtained from one sample of same population in two different time and conditions and each individual gives a pair of observations. Such conditions are commonly faced in biological experiments including field studies when it becomes essential to compare the means of two paired data.

To know significance of difference between two correlated means same formula of t is used here

$$t = \frac{\overline{X}_1 - \overline{X}_2}{SE_D}.$$

But here different formula to obtain standard error of difference is used

$$SE_D = \sqrt{SE\overline{X}_1^2 + SE\overline{X}_2^2 - 2r_{12} \,.\, SE\overline{X}_1 \,.\, SE \,.\, M_2}$$

Here $SE\overline{X}_1$ and $SE\overline{X}_2$ = SE of the first and second mean,

r_{12} = correlation co-efficient between scores made on first and second test.

Example. The weight of 10 pigs when brought in piggery and after six months are given below. Examine whether the gain in weight is statistically significant or not.

Table 6.4.

Weight when brought first	*Weight after six months*
49	52
41	43
37	46
41	52
42	46
37	38
39	42
38	41
41	42
35	38

Calculation. Table of six columns is prepared with the help of above data :

Table 6.5.

Weight when brought first X_1	$X_1 - \overline{X}_1 = x_1$	x_1^2	*Weight after 6 months* X_2	$X_2 - \overline{X}_2 = x_2$	x_2^2
49	9	81	52	8	64
41	1	1	43	−1	1
37	3	9	46	2	4
41	1	1	52	8	64
42	2	4	46	2	4
37	3	9	38	6	36
39	1	1	42	2	4
38	2	4	41	−3	9
41	1	1	42	−2	4
35	5	25	38	−6	36

	On Entry	*after 6 months*
No. of pigs	10	10
Mean ($\overline{X}_1$)	40	44
Standard deviation (σ)	3.88	5
Standard error of mean ($SE\overline{X}$)	1.22	1.58

Difference between means = –4

Correlation between first and last test = 0.50

$$SE_D = \sqrt{(1.22)^2 + (1.58)^2 - 2 \times 0.50 \times 1.22 \times 1.58}$$

$$= \sqrt{1.48 + 2.49 - 1.92}$$

$$= \sqrt{3.97 - 1.92}$$

$$= \sqrt{2.05} = 1.43$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE_D}$$

$$= \frac{-4}{1.43} = -2.79.$$

Here $\quad$ d.f. $= N_1 \; N_2 - 2$

$= 10 + 10 - 2 = 18.$

Significance. On perusal of *t* table at d.f. 18 on .01 level the value of *t* is 2.552. The calculated value of *t* is –2.79, which is higher than the table value. Hence the mean gain in weight is statistically significant.

t-test from grouped data

From grouped data value of *t* is obtained by the same formula but deduction for mean and σ slightly varies. We can see the difference while calculating *t* test of following example :

Example. In an experiment find out the effect of a hormone spray on the seed yield of french beans. Following results are obtained. Draw inference using '*t*' test as regards to the hormonal spray on the seed yield.

X_1 control : 30, 35 31, 36, 38, 32, 25, 39, 31, 34, 33, 35, 40, 30, 32, 28, 26, 29, 30, 30, 35, 36, 37, 38, 39, 30, 31, 39, 40, 35, 36, 26, 27, 40, 41, 35, 38, 32, 31, 38, 30, 35, 36, 25, 28, 29, 37, 36, 36, 30.

X_2 treated : 35, 38, 40, 45, 40, 42, 54, 55, 43, 45, 50, 44, 51, 52, 53, 52, 48, 50, 49, 47, 49, 50, 49, 51, 52, 50, 51, 52, 40, 46, 48, 49, 50, 35, 38, 45, 47, 45, 46, 50, 51, 46, 41, 42, 48, 49, 51, 52, 50, 53.

The values (number of seeds/plant) are arranged in ascending order of magnitude with their frequency values.

Number of seeds/plant (X_1 variables stands for control stock and X_2 for treated stock).

Table 6.6.

Control					Treated				
X_1	f_1	X_1^2	$f_1.X_1^2$	$f_1.X_1$	X_2	f_2	X_2^2	$f_2.X_2^2$	$f_2.X_2$
25	2	625	1250	50	35	2	1225	2450	70
26	2	676	1352	52	38	2	1444	288	76
27	1	729	729	27	40	3	1600	4800	120
28	2	784	1568	56	41	1	1681	1681	41
29	2	841	1682	58	42	2	1764	3528	84
30	7	900	6300	210	43	1	1849	1849	43
31	4	961	3844	124	44	1	1936	1936	44
32	3	1024	3072	96	45	4	2025	8100	180
33	1	1089	1089	33	46	3	2116	6348	138
34	1	1156	1156	34	47	2	2209	4418	94
35	6	1225	7350	210	48	3	2304	6912	144
36	6	1296	7776	216	49	5	2401	12005	245
37	2	1369	2738	74	50	7	2500	17500	350
38	4	1444	5776	152	51	5	2601	13005	255
39	3	1521	4563	117	52	5	2704	13520	260
40	3	1600	4800	120	53	2	2809	5618	106
41	1	1681	1681	41	54	1	2916	2916	54
					55	1	3025	3025	55
	$\Sigma f_1 = 50$		$\Sigma f_1 X_1^2 = 66726$	$\Sigma f_1 X_1 = 1670$		$\Sigma f_2 = 50$		$\Sigma f_2.X_2^2 = 112499$	$\Sigma f_2 X_2 = 2359$

Here, $$\overline{X}_1 = \frac{\Sigma f_1 . X_1}{\Sigma f_1} = \frac{1670}{50} = 33.4$$

$$X_2 = \frac{\Sigma f_2 . X_2}{\Sigma f_2} = \frac{2359}{50} = 47.18$$

$\therefore$ SD of X_1 or $$\sigma X_1 = \sqrt{\frac{\Sigma f_1 . X_1^2 - \frac{(\Sigma f_1 . X_1)^2}{\Sigma f_1}}{\Sigma f_1 - 1}}$$

or $$\sigma^2 X_1 = \frac{\Sigma f_1 . X_1^2 - \frac{(\Sigma f_1 . X_1)^2}{\Sigma f_1}}{\Sigma f_1 - 1}$$

$$= \frac{56726 - \frac{(1670)^2}{50}}{50 - 1}$$

$$= \frac{56726 - 55778}{49}$$

$$= \frac{984}{49} = 19.347$$

$$\sigma X_2 = \sqrt{\frac{\Sigma f_2 . X_2^2 - \frac{(\Sigma f_2 . X_2)^2}{\Sigma f_2}}{\Sigma f_2 - 1}}$$

$$\sigma^2 X_2 = \frac{\Sigma f_2 . X_2^2 - \frac{(\Sigma f_2 . X_2)^2}{\Sigma f_2}}{\Sigma f_2 - 1}$$

$$\sigma^2 X_2 = \frac{112499 - \frac{(2359)^2}{50}}{50 - 1}$$

$$= \frac{112499 - 11297.62}{49}$$

$$= \frac{1201.38}{49} = 24.518.$$

Pooled variance[2] of X_1 and X_2

$$= \frac{19.347}{50} + \frac{24.518}{50}$$

$$= 0.387 + 0.49 = 0.877.$$

Pooled standard error of difference

$$\sigma d = \sqrt{\sigma_d^2} = \sqrt{0.877} = 0.936$$

$$\overline{X}_1 - \overline{X}_2 = 33.4 - 47.28$$

$$= 13.88$$

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_d} = \frac{13.88}{0.936} = 14.829.$$

Here d.f. $= N_1 + N_2 - 2$
$= 50 + 50 - 2$
$= 98.$

Calculated value of t comes to 14.829.

Table value of t on df_{98} at 5.0% level is 1.96.

Inference. The calculated value of 't' 14.829 is very high than the tabulated 't' value 1.96. Therefore, we can say that the hormone spray definitely play significant effect on the seed yield.

EXERCISE

1. What do you mean by student's 't' test ? Mention formula to obtain 't' ratio.
2. What is degree of freedom ? Mention formula to obtain d.f. in single and paired data while computing 't' ratio. If we get calculated 't' value higher than the table 't' value at a certain d.f. and level then the experiment is reliable or not.
3. In a pig form of military camp at Gaya a group of 12 pigs were fed with one variety of food A while another group of 12 pigs of same age, sex and stock were given another variety of food B. After one month weight of both groups was recorded as follows :
   First group on food A : 31, 34, 34, 29, 26, 32, 35, 38, 34, 30.
   Second group on food B : 26, 24, 28, 29, 30, 29, 32, 26, 35, 29.
   As certain whether the difference in mean weights of First and Second group (due to different food quality) is significant or not. For your answer use students 't' test.
4. Two samples X_1 and X_2 have means 2.158 and 1.96 respectively. Determine whether there is significant difference in *mean* of sample A and sample B.
   Sample A : 2.5, 2.1, 1.9, 2.4, 1.8, 2.6, 2.5, 2.3, 0.9, 1.7, 2.2, 2.0.
   Sample B : 1.8, 1.7, 1.9, 2.2, 2.3, 2.0, 1.9, 1.8, 2.1, 1.9.
5. In a study on growth of children, one group of 100 children had a mean height of 50 cm and standard deviation of 2.5 cm while another group of 150 children had a mean height of 62 cm and standard deviation of 3 cm. Is the difference between the two statistically significant ?
6. Average length of fishes in a pond is 175 cm. 10% are more than 180 cm. If the length distribution of fishes is normal then find out the standard deviation. Number of soldiers in pond is 500.

7. The Chi-Square Test

> **Synopsis.** *Introduction, Definition, Common applications of χ^2, Pre-requisite pf χ^2 test, Methods to draw inferences, calculation of χ^2 test. Exercise.*

Introduction. Standard error and student's *t*-test are parametric test and are applied to only quantitative data. In biological experiments a non-parametric test is very commonly used called *Chi-square* test. It is applied only for qualitative data such as colour, health, intelligence, cure response of drug etc. which do not have numerical values.

Chi-square test was developed by Prof. A.R. Fisher in 1870. Karl Pearson improved Fisher's chi-square test in its modern form in 1900. Chi-squre is derived from the Greek letter (Chi χ) and pronounced as 'Kye'.

Definition. "*Chi-square test is the test of significance of overall deviation square in the observed and expected frequencies divided by expected frequencies*".

General computing formula for Chi-square

$$\chi^2 = \Sigma \left\{\frac{(f_0 - f_e)^2}{f_e}\right\} \quad \text{or} \quad \Sigma\left\{\frac{(0 - E)^2}{E}\right\}.$$

Here f_0 or 0 = Observed frequency
f_e or E = Expected frequency.

Common applications of χ^2 test :

1. As an alternate test to find significance of difference in two or more than two proportions. Chi-square test is a very useful test which is applied to find significance in the same type of data with two more advantages :

(*a*) To compare the values of two binomial samples even if they are small such as oxygen consumption in 5 control and 5 thyroxine (T_4 – 1 mg) injected fishes of same species.

(*b*) To compare the frequencies of two multinomial samples such as oxygen consumption in control and T_4 injected groups of fishes weighing 1–10 g, 11–20 g, 21–30 g, 31–40 g and 41–50 g.

2. As a test of association between two events in binomial or multinomial samples. Chi-square measures the probability of association between two discrete attributes. Two events can be studied for their association such as iron intake and Hb%, season and fecundity, T_4 injection

and oxygen consumption, nutrition and intelligence, weight and diabetes etc. There are two possibilities, either they influence or they do not. χ^2 is very useful tool in ascertaining Mendilian ratio.

Association table : Table is prepared by enumeration of qualitative data. Since one wants to know the association between two sets of events, the table is also called association table.

Four fold or 2 × 2 contingency table : When there are only two samples, each divided into two classes, it is called *fourfold,* four cell of 2 × 2 *contingency table* (Table 7.1).

Table 7.1. Outcome of treatment with drug and placebo.

Groups	*Outcome or Result*		
	Died	*Survived*	*Total*
A (Control on placebo)	10	25	35
B (Experiment, on drug)	5	60	65
Total	15	85	100

Multifold Association Table : The association of two sets of events having more than two classes will be larger than fourfold or four cell contingency table (Table 7.2).

Table 7.2. Social class and Wuchereria positivity

Social class	*Outcome or Result*			
	Number +ve	*Number −ve*	*Total*	*Percentage +ve*
I.	4	76	80	5
II.	20	180	200	10
III.	60	440	500	12
IV.	144	576	720	20
	228	1272	1500	47

3. As a test of goodness of fit. *Chi-square* test is also applied as a test of "goodness of fit". Goodness of fit reveals the closeness of observed frequency with those of the expected. Thus it helps to answer whether something (physical or chemical factors) did or did not have an effect. If observed and expected frequency are in complete agreement with each other then the *chi-square value* will be zero. But it rarely happen in biological experiments. There is always some degree of deviation.

Pre-requisites of χ^2 test. There are three basic pre-requisites of χ^2 test such (*i*) Sample must be random. (*ii*) Data should be qualitative. (*iii*) Observed frequency should not be less than five.

Method to draw inferences. If the calculated value of χ^2 is less than the tabulated value of χ^2, then observed and expected value is considered insignificant (appendix 4). But if the calculated value of χ^2 is more than the tabulated value then the two variables are dependent on each other and value is significant.

The quantity in the denominator which is one less than the independent number of observation in a sample is called degree of freedom. If there are 2 classes (For example **control** and T_4 injected, male and female) the degree of freedom would be $2 - 1 = 1$. If there are three classes then $d.f = 3 - 1 = 2$, in case of 4 classes $d.f. = 4 - 1 = 3$ and so on.

If the χ^2 value obtained in more than two pairs of data then $d.f = (2 - 1) \times (2 - 1) = 1$.

Calculation of χ^2 test. The calculation of χ^2 is easy and same for each case. If f_0 is the observed frequency of a particular category of a variable and f_e is the expected frequency of some hypothesis, then χ^2 is calculated by following formula

$$\chi^2 = \Sigma \left\{ \frac{(f_0 - f_e)^2}{f_e} \right\}.$$

Following steps have to be taken to obtain χ^2 value :

— Make a contingency table and note the observed frequency (f_0 or 0) in each class of one event, row-wise *i.e.* horizontally and then the numbers in each group of the other event, columnwise *i.e.* vertically.

— Determine the expected frequency (f_e or E) in each group of the sample on the assumption of null hypothesis (H_0) *i.e.* no difference in the proportion of the group from that of the population.

— Find the difference between the observed and expected frequency in each cell ($f_0 - f_e$) or (0 − E).

— Calculate the χ^2 value applying formula :

$$\chi^2 = \Sigma \left\{ \frac{(f_0 - f_e)^2}{f_e} \right\}.$$

— Calculated χ^2 value is compared with tabulated value of χ^2 at desired degree of freedom under different probabilities 0.5, 0.1, .05, .01, .001 etc.

— If calculated χ^2 value is higher than tabulated value then it is considered as significant. But if calculated value is less than the table value then it is considered as insignificant.

Example. In a grassland the earthworm population was sampled from ten randomly located plots of 1 m^2 area. The following table gives the number of earthworms obtained. Examine the distribution pattern of earthworms :

Area	1	2	3	4	5	6	7	8	9	10
Earthworm No./m²	25	32	17	23	15	39	27	19	22	26

$$\Sigma X = 245\ ; \quad \text{Mean} = \frac{\Sigma X}{N} = \frac{245}{10} = 24.5$$

The expected number of earthworms is determined taking into consideration that the population is equally distributed in all quadrates. Thus the expected number of earthworms in each quadrant is the mean number of earthworms. Following table gives the summary of the results :

Table 7.3.

Observed	25	32	17	23	15	39	27	19	22	26
Expected	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5
Difference	0.5	7.5	−7.5	−1.5	−9.5	14.5	2.5	−5.5	−2.5	1.5

$$\chi^2 = \frac{(0.5)^2}{24.5} + \frac{(7.5)^2}{24.5} + \frac{(-7.5)^2}{24.5} + \frac{(-1.5)^2}{24.5} + \frac{(-9.5)^2}{24.5}$$

$$+ \frac{(14.5)^2}{24.5} + \frac{(2.5)^2}{24.5} + \frac{(5.5)^2}{24.5} + \frac{(-2.5)^2}{24.5} + \frac{(1.5)^2}{24.5}$$

$$= .01 + 2.29 + 2.29 + .09 + 3.6 + 8.58 + .25 + 1.23 + .25 + .009$$

$$= 18.68.$$

Significance. Here $d.f. = 10 - 1 = 9$.

The tabulated value of χ^2 at d.f. 9 on $p = 0.05$ is 16.92 and the calculated value is higher *i.e.* 18.68. This shows that the two series of frequencies, observed and expected are different, indicating that the earthworms population is not distributed equally.

Example. In a monohybrid cross between tall (TT) and dwarf (*tt*) 1574 tall and 554 dwarf were obtained. Suggest if a ratio of 3 : 1 is suitable or not.

Calculation. Total number = 1574 Tall + 554 Dwarf
= 2128.

Therefore expected 3 : 1 will be $2128 \times \frac{3}{4} : 2128 \times \frac{1}{4}$.

Expected ratio = 1596 : 532
Observed ratio = 1574 : 554.

Putting the values in the formula :

$$\chi^2 = \Sigma\left\{\frac{(f_0 - f_2)^2}{f_e}\right\}$$

$$= \frac{(1574-1596)^2}{1596} + \frac{(554-532)^2}{532}$$

$$= \frac{(-22)^2}{1596} + \frac{(22)^2}{532} = \frac{484}{1596} + \frac{484}{532}$$

$$= 0.303 + 0.909 = 1.212. \quad \textbf{Ans.}$$

Here, d.f. = 2 – 1 = 1.

Significance. At 5% level, at 1 degree of freedom the table value of $\chi^2 = 3.84$ and the calculated value of χ^2 is 1.212.

This shows that the two series of frequencies, observed and expected is not in agreement with the theoretical ratio of 3 : 1.

Example 3. In a cross between black male and gray female Drosophila the offspring obtained were 25 black and 35 gray. Calculate the χ^2 and give your inference on the ratio of black and gray offsprings. Expected number is calculated from the fact that gray body colour is dominant and the expected ratio of this nature is 1 : 1. [Total number of offsprings are 60].

Calculation. Following table is prepared to obtain required values.

Table 7.4.

Black	*Gray*
Observed number 25	35
Expected number 30	30
(0 – E) 25 – 30 = –5	35 – 30 = +5
$(0 - E)^2$ $(25 - 30)^2 = 25$	$(35 - 30)^2 = 25$

Putting values in the formula :

$$\chi^2 = \Sigma\left\{\frac{(f_0 - f_e)^2}{f_e}\right\} = \frac{25}{30} + \frac{25}{30}$$

$$= \frac{5}{6} + \frac{5}{6} = \frac{10}{6} = 1.66.$$

The table value of χ^2 at d.f. 1 p 0.05 is 3.84. The obtained value is 1.6. The obtained value is less and therefore it is not in agreement with the theoretical ratio of 1 : 1.

Example 4. RBCs count lac/mm^3 and Hb% gm/100 ml of 500 persons of a locality were recorded as follows. Is there any significant relation between RBC count and Hb%. Find it by χ^2 method.

RBCs count	*Hb%*		
	Above normal	*Below normal*	*Total*
Above normal	85	76	160
Below normal	165	175	340
Total	250	250	500

Calculation. Make following two tables to obtain χ^2

Table 7.5.

RBCs count	*Hb%*		
	Above normal	*Below normal*	*Total*
Above normal	$O = 85$ $E = \frac{250 \times 160}{500} = 80$	$O = 75$ $E = \frac{250 \times 160}{500} = 80$	160
Below normal	$O = 165$ $E = \frac{250 \times 340}{500} = 170$	$O = 175$ $E = \frac{250 \times 340}{500} = 170$	340
Total	250	250	500

On the basis of above data following table is prepared :

Table 7.6.

O	E	$O - E$	$(O - E)^2$	$\frac{(O-E)^2}{E}$
85	80	$85 - 80 = 5$	25	$\frac{25}{80} = 0.31$
165	170	$165 - 170 = -5$	25	$\frac{25}{170} = 0.14$
75	80	$75 - 80 = -5$	25	$\frac{25}{80} = 0.31$
175	170	$175 - 170 = 5$	25	$\frac{25}{170} = 0.15$
				$\Sigma \frac{(O-E)^2}{E} = 0.90$

Here $d.f. = (2-1) \times (2-1) = 1.$

Significance. At 5% level, on 1 d.f. the table value of $\chi^2 = 3.84$. The calculated value is very less *i.e.*, 0.90. It indicates that Hb% and RBC count are independent to each other.

[**Note :** In fact Hb% and RBCs count are not independent to each other.]

This may be due to hypothetical data.

EXERCISE

1. Define and mention formula of χ^2 test. What do you mean by goodness of fit ?
2. Determine if there is any association between whooping cough and tonsillectomy. When in a random sample of 100 children of a school, 25 had history of tonsillectomy and 60 of whooping cough and 10 had both while 25 had none. [Ans. 5.55]
3. Three groups with 20 patients in each were administered analgesics A, B and C. Relief was noted in 20, 10, and 6 cases respectively. Is this difference due to the drug or by chance ? [Ans. 8.6]
4. In an experiment the effect of two concentrations of a pesticide "Dimercron–100" was studied in relation to mortality of the fish, Tilapia mossambica, in two different aquaria. The results were noted after 24 hours of treatment. It was observed that some fish in both the aquaria died and the percentage of death was 27% in the lower concentration and 39% in the higher concentration. It is to test whether the percentage of death in the higher concentration is significantly different from that of lower concentration or both are independent. [Ans. 2.812]
5. The survey report of a hospital showed that out of 1000 patients 432 were men and 568 were women. The report further revealed that 305 of men and 355 of the women suffered from high blood pressure. Test the hypothesis that blood pressure was equally frequent in men and women using a 2×2 contingency table. [Ans. $\chi^2 = 6.82$]
6. During census one surveyor obtained the following results as regards the male and female population of a village. Do the results agree with the theory that the male and female populations are in 1 : 1 ratio ?
   Male population = 968.
   Female population = 1042.

8. Probability

Synopsis. *Introduction, Terminology, Definition of Probability, Calculation of Probability of simple events. Rules of Probability—Addition theory and Multiplication theory ; Random variable and probability distribution, Theoretical probability distributions.*

Introduction. Man is generally interested in drawing conclusions from events which involve uncertainties. For example, the tossing of a coin, birth of a male or female child etc. involves an act that completely depends on pure chance. Measures of chances expressed numerically is the subject matter of probability theory.

The probability theory provides a means of getting an idea of the likelihood of occurrence of different events resulting from a random experiment in terms of quantitative measures ranging between zero and one. The probability is zero ($p = 0$) for an impossible event and one for an event which is certain to occur ($p = 1$). Chances of survival after rabies infection is impossible that is $p = 0$. The death of living being is inevitable event *i.e.* $p = 1$. The other degrees of uncertainties or the likelihood of occurrence of events are indicated by probabilities ranging between zero and one.

If 'a' stands for an event to happen and 'b' stands for an event not to happen then the probability of its happening 'p' can be symbolized as follows :

$$p = \frac{a}{a+b}.$$

The probability of failure of event 'q' can be symbolized as follows

$$q = \frac{b}{a+b}.$$

Arithmetically probability 'p' of a positive event may be calculated with the help of following formula :

$$p = \frac{\text{Number of cases to happen}}{\text{Number of cases including failure}}$$

'p' and 'q' added together will be equal to 1.

Symbolically $p + q = 1$. If $p = 0.4$ then 'q' will be

$$1 - p\ (1 - 0.4 = 0.6).$$

Terminology. For an understanding of the concept of probability theory, the following terminology must be clearly understood :

1. **Sample Space.** The set of results obtained from an experiment is called sample space or probability space. Its symbol is S or Ω and the number of components of this set is denoted by n(S).

For example, (*i*) A woman can give birth either to a male child or a female child. In this context sample space or probability space, set is {M, F}.

(*ii*) Suppose a bitch can produce maximum 7 child at a time. Then the probability of giving birth at one time will be any of the seven and sample space set will be

$$\{1, 2, 3, 4, 5, 6, 7\}.$$

(*iii*) If two women giving birth then sample space = {(M, M), (M, F), (F, M), (F, M)}. Here (M, M) means both mothers giving birth to male child. (M, F) means first lady giving birth to male and 2nd lady giving birth to female child. (F, M) means first lady giving birth to female child and second lady giving birth to male child. (F, F) means both women ladies birth to female child. It is written as follows :

S = {M, F} × {M, F} or {MM, MF, FM and FF} and sample number of set n(S) = 4.

In context to above example sample space may be prepared by a simple method :

Woman I → *II* ↓	*M*	*F*
M F	MM FM	MF FF

2. **Sample point.** Components of sample space is known as sample point. Suppose S = {1, 2, 3, 4, 5, 6, 7}, then 1, 2, 3, 4, 5, 6, 7 are sample points. In case of S = {M, F} M and F are sample points.

3. **Trial and Events.** An experiment is called a trial and all its possible outcomes are known as events. For example, tossing of a coin is an experiment, and one toss of the coin will constitute a trial. The experiment of tossing a coin has only two outcomes or events, head or tail. Example of another experiment may be throw of a die. One throw of the die constitute a trial. There are six possible events, *i.e.*, the turning up of any of the six numbers , 2, 3, 4, 5 or 6 in one trial. To get even number is an event (E_1) and to get odd number is another event (E_2). To get bigger number than 4 is event (E_3) and to get smaller number than 5 is event (E_4). Even number

in sample space is 2, 4, 6 therefore event $E_1 = \{2, 4, 6\}$. Odd number in sample space is 1, 3, 5, therefore event $E_2 = \{1, 3, 5\}$. Likewise $E_3 = \{5, 6\}$ and $E_4 = \{1, 2, 3, 4\}$.

There are various types of events. Few are as follows :

(*i*) **Simple events.** Occurrence or non-occurrence of single event *i.e.*, a simple event cannot be decomposed into a combination of other events. Drawing of a particular card from a pack is a simple event. A woman giving birth to a child — S = {M, F} ; {M} and {F} is simple event. Two women giving birth to child S = {MM, MF, FM, FF} ; {MM}, {MF}, {FM}, {FF} are simple events.

(*ii*) **Mixed or compound or joint events.** Occurrence of two or more simple events simultaneously is called mixed events. For example, if a bag contains 4 white and 6 red balls and we make draw of 2 balls at random, then the events that 'both are white' or 'one is white' and 'one is red' are compound events. Similarly, if two persons X and Y draw one card each from a pack of cards simultaneously, such results as 'both cards are king' are compound events. Compound events may be of two types :

(*a*) **Independent events.** If the occurrence of one event does not affect the occurrence of another, they are said to be independent events. For example, if a coin is tossed twice, the result of the second toss would in no way be affected by the result of the first toss.

(*b*) **Dependent events.** If the occurrence of one event influences the occurrence of the other, then the second event is said to be dependent on the first. For example, if a person draws a card from a full pack and does not replace it, the result of the draw made afterwards will be dependent on the first draw.

(*iii*) **Equally likely event.** If the likelihood of the occurrence of every event is the same it is called equally likely event. For example, in birth of a child, the male and female have an equal chance. Similarly in a throw of die, each one of the faces marked 1, 2, 3, 4, 5, 6 has an equal chance of coming on top.

(*iv*) **Mutually exclusive events.** When occurrence of one event implies that the other cannot occur then two events are called mutually exclusive. For example, in a birth of a child either male or female is born *i.e.*, the two events, birth of male and female both cannot occur simultaneously (birth of twins is exceptional to rule).

(*v*) **Sure event.** Every set is sub-set of self event therefore sample space is also an event. It is said to be sure event and denoted by the same symbol of sample space or Ω.

(*vi*) **Null or impossible events.** No chance of getting that event is said to be impossible event. It is denoted by ϕ. Probability of getting 7 in a throw of die is a null event because die do not have sample point 7.

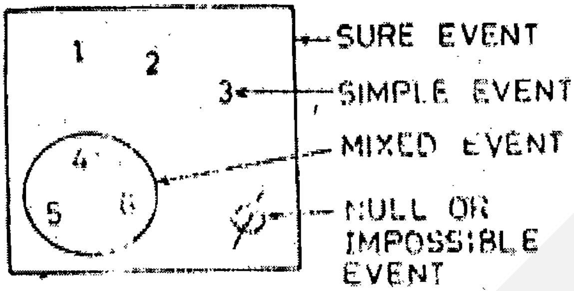


Fig. 8.1. A die exhibiting different events.

Definition of Probability : There are two definitions of probability :

(1) 'Classical' or mathematical.

(2) 'Empirical' or statistical.

(1) **Classical definition.** If the trial of an experiment results in n exhaustive, mutually exclusive and equally likely cases and m of them are favourable to the happening of an event E, then the probability of the event E may be given by the following formula :

$$P(E) = \frac{m}{n} = \frac{\text{Favourable number of cases}}{\text{Exhaustive number of case}}$$

and the probability that the event does not happen.

$$P(\bar{E}) = \frac{n-m}{n} = \frac{\text{No. of cases unfavourable to event E}}{\text{Exhaustive number of cases}}$$

In this way we observer that

$$P(E) + P(\bar{E}) = \frac{m}{n} + \frac{n-m}{n} = 1$$

or $\quad P(E) = 1 - P(\bar{E})$

and $\quad P(\bar{E}) = 1 - P(E).$

It is now clear from the above definition that :

(*a*) The probability of an event is the ratio of the number of favourable cases to the exhaustive number of cases in a trial.

(*b*) The probability of an event which is sure to happen is 1.

(*c*) The probability of an event which cannot occur is 0.

(*d*) The sum of the probabilities of happening and not happening of an event is always equal to 1.

When the various outcomes of a trial are not equally likely the classical definition of probability fails to give the probability of an event. If the exhaustive number of cases (n) in a trial is infinite, then also the definition fails to give the required probability.

Example. In a throw of die, S = {1, 2, 3, 4, 5, 6} and getting even number {2, 4, 6}, then $n(S) = 6$, $n(E) = 3$

and $$P(E) = \frac{n(E)}{n(S)} = \frac{3}{6} = \frac{1}{2}.$$

(2) **Empirical or statistical definition.** If a trial is repeated a number of times (n times) in same condition and obtained event is p times then the limiting value of the ratio of the number of times the event E happens to the number of trials, is called the probability of the event E.

Symbolically, if in n trials, an event E happens pn times then the probability of E is given by

$$P(E) = \underset{n \to \infty}{\text{Lt}} \frac{P_n}{n} = P$$

p_1 is the limit value of relative frequency of E which happens during the experiment with relative frequency. In simple words one can say that "Probability measures the relative frequency of a particular event happening by chance in long run."

Example. The reproduction results of a woman may reveal that she has given birth to only sons or daughters though the probability of getting both were 50%. If the range of observation is extended upto population then the ratio of male and female child birth comes to 1 : 1. It proves that by increasing the number of observation the value of m/n becomes nearer to actual probability and that is why we use $n \infty$. Ratio P_n/n is called relative frequency. In this case there is no condition like equally likely event.

The probability of death by heart attack or diabetes etc. in a year of any country or globe may be obtained by the ratio of death number and population p_n/n.

Calculation of probability of simple events

Example 1. A woman gives birth to a child. What is the probability of getting male child ?

Sol. In giving birth to a child, there may be two outcomes, *i.e.*, male or female, thus the exhaustive number of cases = 2.

Now, one of them, *i.e.*, the birth of male child is a favourable event, therefore, the probability of the required event

$$P(M) = \frac{\text{Favourable number of cases}}{\text{Exhaustive number of cases}} = \frac{1}{2}.$$

Example 2. Two cards are drawn from a well-shuffled pack. Find the probability that :

(*i*) **both are kings**
(*ii*) **one king and one queen**
(*iii*) **both are spades.**

Sol. (*i*) **Probability of 2 kings.** Let E_1 denotes the event that both cards are kings. Since, out of 52 cards 2 can be drawn in $52C_2 = \frac{52 \times 51}{1 \times 2}$ = 1326 exhaustive number of cases. There are four kings in the pack and out of these 2 can be selected in $4C_2 = \frac{4 \times 3}{1 \times 2} = 6$ cases which are favourable to desired event E_1.

Hence, $$P(E_1) = \frac{6}{1326} = \frac{1}{221}.$$

(*ii*) **Probability of one King and one Queen.** Let E_2 denotes the event of getting one King and one Queen. These are four Kings and 4 Queens in a pack. Therefore, favourable ways of drawing a King and a Queen simultaneously $= 4 \times 4 = 16$. Hence,

$$P(E_2) = \frac{16}{1326} = \frac{8}{663}.$$

(*iii*) **Probability of both cards being spades.** Let E_3 denotes the event that both cards are spades.

There are 13 spades and out of these 2 can be selected in

$$^{13}C_2 = \frac{13 \times 12}{1 \times 2} = \frac{156}{2} = 78.$$

favourable numbers of cases.

Therefore, $$P(E) = \frac{78}{1326} = \frac{1}{17}$$

Example 3. 5000 fishes were brought from a pond in a laboratory. Average age of these fishes are 800 hour and standard deviation is 150 hour. Find probable death of fishes before 600 hour.

Sol. Here $N = 5000$, $\mu = 800$ hour (mean of population), $\sigma = 150$ hour.

Probability of death between 0 – 600 hour.

Here $$Z = \frac{\overline{X} - \mu}{\sigma n}$$

$$= \frac{600 - 800}{150}$$

$$= \frac{-200}{150} = -1.33.$$

Now see Z area = 1.33 in normal curve table between area 0 and 600 = 0.5 – 0.4082 = 0.0918.

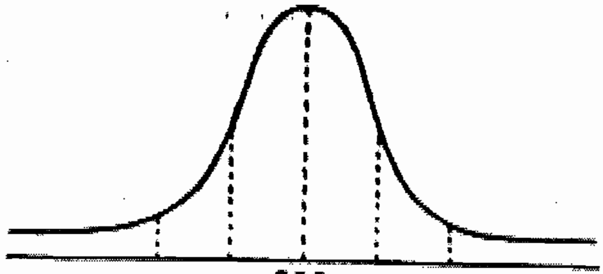


Fig. 8.2. Normal curve showing average age 800 hour.

Probability of death between 0 – 600 is 9.18% fishes.

$\therefore$ No. of dead fishes between 0 – 600 hour

$$= .0198 \times 5000$$
$$= 459 \text{ fishes. } \textbf{Ans.}$$

Rules of probability. There are two common rules of calculating probabilities which are useful in simplifying the procedure of calculating probabilities of *mutually exclusive* and *compound events.*

(1) **Addition rule of probability** (when events are mutually exclusive): When two events, say E_1 and E_2 are mutually exclusive (Events can not occur simultaneously) the probability of occurrence of either E_1 and E_2 is the sum of the probabilities of the individual events.

Symbolically, if $P(E_1)$ and $P(E_2)$ are the respective probabilities of two mutually exclusive events E_1 and E_2, then the probability that one of them happens is given below :

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2).$$

The rule can be extended to any number of mutually exclusive events as below :

$$P(E_1 \text{ or } E_2 \text{ or } E_3 \ldots \text{ or } E_x) = P(E_1) + P(E_2) + P(E_3) + \ldots P(E_x)$$

Example. Of every 100 students who are selected, we find, on the average :

10 students were in grade E_1.
25 students were in grade E_2.

Use addition rule to find the probability that the selected student will have either grade A or B.

Sol. Let E_1 and E_2 be the events that the selected student has grade A and B respectively. Therefore,

$$P(E_1) = \frac{10}{100} = 0.10$$

$$P(E_2) = \frac{25}{100} = 0.25$$

Since both the events E_1 and E_2 are mutually exclusive, the probability that any one of them happens can be obtained by using addition rule.

$$P(A \text{ or } B) = P(E_1) + P(E_2)$$
$$= 0.10 + 0.25 = 0.35.$$

(2) (*a*) **Multiplicative rule of probability** (when events are independent) : The probability of two or more independent events occurring together is the product of the probabilities of the individual events.

Symbolically, if $P(E_1)$ and $P(E_2)$ are the respective probabilities of happening of two independent events E_1 and E_2, then the probability that the two events will happen together is given below :

$$P(E_1 \text{ and } E_2) = P(E_1) \,.\, P(E_2).$$

This rule can be extended to any number of independent events E_1, E_2, E_3 ... E_x as below :

$$P(E_1 \text{ and } E_2 \text{ and } E_3 \text{ ...., and } E_x) = P(E_1).P(E_2).P(E_3) \text{ ... } P(E_x).$$

Example. When two children are born one after the other, the possible sequences will be any of the following four :

	Sequence	*Probability*
(1)	M and M	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
(2)	M and F	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
(3)	F and M	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
(4)	F and F	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

Chance of getting two male child $= \frac{1}{4} = 25\%$

Chances of getting two female child $= \frac{1}{4} = 25\%$

Chances of getting one of the either sex will be total of second and third sequence $= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ so, if a female child is born first the probability of the second issue being male will be 75% and its being female 25%.

The probability of sequences (2) and (3)

$$= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}$$

$$= \frac{1}{4} + \frac{1}{4} = \frac{2}{4} = \frac{1}{2} = 50\%.$$

WEB: www.evolution.co.in | MAIL: info@evolution.co.in | PHONE: +91 11 4709 2329 ; + 91 9810 975 937 | ©Evolution Educare Pvt. Ltd.

Therefore probability of two female child = 25% so that of second child being male = 1 – 25% = 75%.

(2) (*b*) **Multiplicative rule** (when events are dependent) : Before dealing multiplicative rule one should know about the concept of conditional probability.

Conditional probability. If the events E_1 and E_2 are dependent so that the probability of occurrence of E_2 is affected by the occurrence of E_1. Then the probability of an event E_2 occurring when it is known that an event E_1 occurred is called the conditional probability and is denoted by $P(E_2/E_1)$. The term $P(E_2/E_1)$ may be read, "The probability of occurrence of E_2, given that E_1 has already occurred."

Now, the probability that both dependent events E_1 and E_2 occur in that order is the probability that E_1 occurs multiplied by the conditional probability that E_2 occurs given that E_1 has already occurred. Symbolically, this multiplicative rule may be written as follows :

$$P(E_1 \text{ and } E_2) = P(E_1) \,.\, P\left(\frac{E_2}{E_1}\right).$$

Example. Three groups of children contain respectively 3 girls and 1 boy ; 2 girls and 2 boys ; 1 girl and 3 boys. One child is selected at random from each group. Find the probability that the three selected children include 1 girl and 2 boys.

Sol. In given condition, 1 girl and 2 boys may be selected in the following three mutually exclusive events E_1, E_2 and E_3 :

(*i*) Event E_1 – girl from 1st group and boys from 2nd and 3rd groups.

(*ii*) Event E_2 – girl from 2nd group and boys from 1st and 3rd groups.

(*iii*) Event E_3 – girl from 3rd group and boys from 1st and 2nd group.

Each of these events is itself a compound event of three simple independent events. For example, occurrence of event E_1 includes the simultaneous selection of a girl from 1st group, a boy from 2nd group and a boy from 3rd group. Thus, the probability of event E_1 is the multiplication of these events, *i.e.*,

$$P(E_1) = \frac{3}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{9}{32}$$

$$P(E_2) = \frac{1}{4} \times \frac{2}{4} \times \frac{3}{4} = \frac{3}{32}$$

$$P(E_3) = \frac{1}{4} \times \frac{2}{4} \times \frac{1}{4} = \frac{1}{32}$$

Since the three events E_1, E_2 and E_3 are mutually exclusive, therefore the probability that any one of them happens is given below ;

$$P(E_1 \text{ or } E_2 \text{ or } E_3) = P(E_1) + P(E_2) + P(E_3)$$
$$= \frac{9}{32} + \frac{3}{32} + \frac{1}{32} = \frac{13}{32}.$$

Random variable and probability distribution :

A random variable is a numerical quantity obtained by the outcome of a random experiment. In a random experiment, every outcome has a probability and this probability can be assigned to each value of the random variable. For example, if X is a random variable showing the numerical value of the outcome in an observation of throw of a die. Since die has 6 faces 1, 2, 3, 4, 5 and 6, X has six possible outcomes 1, 2, 3, 4, 5, 6 and to each outcome there is an associated probability 1/6 as shown below :

Table 8.1.

Outcome on a die = X	1	2	3	4	5	6	Total
Probability of X = p(X)	1/6	1/6	1/6	1/6	1/6	1/6	1.0

Above table listing all possible outcomes of a random variable X together with corresponding probabilities $p(X)$ is called a probability distribution of X. One thing is notable here that the sum of all probabilities in a probability distribution is 1, *i.e.*, A random variable is either discrete or continuous.

The distributions of discrete and continuous random variables are accordingly called discrete and continuous probability distributions.

Theoretical probability distributions

In previous chapters we have discussed observed frequency distributions which are results of actual observations and experimentation. But sometimes the knowledge of that distribution becomes essential which are not based on real observations. We can deduce frequency distribution of hypothetical values mathematically. All possible hypothetical values of a random variable is known as theoretical probability distribution.

In theoretical probability distributions instead of observed scores there are all possible values of a random variable and the frequencies are replaced by actual probabilities, which depend on the nature of a random variable. In Table 8.1 the probability for every occurrence of 1, 2, 3, 4, 5, 6 is given as 1/6 which we calculate on the basis of a theoretical consideration that all the outcomes are equally likely.

Theoretical probability distribution are of three types :

(1) Binomial probability distribution.
(2) Poisson probability distribution.
(3) Normal probability distribution.

(1) **The binomial probability distribution.** If chances of occurrence of an event, in an experiment is p and no chance to occur that event is q, then in n independent experiments x chance of occurrence of an event is

$$p(x) = {}^n c x p^x q^{n-x}.$$

Probability distribution given by above equation is called Binomial distribution.

Value of x in above equation is 0, 1, 2, 3, ... n, then

$p(1)$ = chances of one success of probability $= {}^n c_1\, p^1 q^{n-1}$
$p(2)$ = chances of two success of probability $= {}^n c_2\, p^2 q^{n-2}$
$p(3)$ = chances of three success of probability $= {}^n c_3\, p^3 q^{n-3}$
$p(r)$ = chances of r success probability $= {}^n c_r\, p^r q^{n-r}$ etc.

It means these probabilities have been taken by the expansion of $(q + p)^n$ in serial steps.

$$(q+p)^n = q^n + {}^n c_1 p q^{n-1} + {}^n c_2 p^2 q^{n-2} + \dots + {}^r c_n p^n$$

If N set is there and in each set there is n attempt then relative frequency can be given by following steps of expansion

$$N(q+p)^n = N[q^n + {}^n c_1 p q^{n-1} + {}^n c_2 p^2 q^{n-2} + \dots + {}^n c_n p^n]$$

where 0, 1, 2, 3, ..., n success is as per set n_0 is Nq^n, $N.{}^n c_1 p q^{n-1}$, $N.{}^n c_2 p^2 q^{n-2}$ ..., $N.{}^n c_n p^n$.

Example 1. In following table r = no. of occurrence of event and f = relative frequency. We have to detect whether distribution is Binomial or not and what is its mean and standard deviation (σ).

Table 8.2.

r	0	1	2	3	4	5	6
f	729	1458	1215	540	135	18	1
	0	1458	2430	1620	540	90	6

$$\Sigma Xf = 6144 \;;\; \Sigma f = 4096$$

Calculation. Here $q^6 \propto 729$ and $q^6 \propto 1$

$$\therefore \quad \frac{p}{q} = \frac{1}{(729)^{1/6}} = \frac{1}{3} \quad \Rightarrow \quad 3p = q = 1 - p$$

$$\Rightarrow \quad 4p = 1 \qquad [\Rightarrow \text{ It implies that}]$$

$$\therefore \quad p = \frac{1}{4} \text{ and } q = \frac{3}{4}.$$

This indicates the distribution is binomial in which

$$p = \frac{1}{4} \text{ and} = \quad q = \frac{3}{4} \quad \text{and} \quad n = 6$$

Therefore mean $= np = 6 \times \frac{1}{4} = \frac{3}{2}$

and standard deviation $= \sqrt{npq} = \sqrt{6 \times \frac{1}{4} \times \frac{3}{4}}$

$$= \frac{3}{2\sqrt{2}} = \frac{3}{2 \times 1.414}$$

$$= \frac{3}{2.828} = 1.06.$$

Mean could also be deduced by following formula

$$\frac{\Sigma X.f}{\Sigma f} = \frac{6144}{4096} = \frac{3}{2} = 1.5. \quad \textbf{Ans.}$$

Example 2. What are the chances of getting any combination *i.e.* 2 boys, 2 girls or one boy and one girl, when number of pregnancies is 2. The probability of these 3 outcomes can be calculated by the formula :

$$(p + q)^2 = p^2 + q^2 + 2pq$$

p^2 is probability of getting 2 boys, q^2 of 2 girls and $2pq$ of one boy and one girl.

[By observing large number of births in a universe chances of male birth may be 51% and female birth 49% so $p = 0.51$ and $q(1 - 0.51) = 0.49$].

Substitute the probability p and q in the above formula *i.e.* 0.51 and 0.49.

$$\begin{aligned}(p + q)^2 &= (0.51)^2 + (0.49)^2 + (2 \times 0.51) \times 0.49 \\ &= 0.2601 + 0.2401 + 0.4998.\end{aligned}$$

Probability of proportional chances of getting 2 boys are 0.2601 or 26.01% of getting 2 girls are 0.2401 or 24.01% and of getting one boy and one girl are 0.4998 or 49.98%.

The same may be extended to birth of 3 children, possible combinations will be 3 boys (p^3) ; 3 girls (q^3) ; one boy and 2 girls ($3pq^2$) ; 2 boys and one girl ($3p^2q$). To calculate the percentage chances of these combinations apply the formula :

$$\begin{aligned}(p + q)^3 &= p^3 + q^3 + 3pq^2 + 3p^2q \\ &= 0.51^3 + 0.49^3 + 3 \times 0.51 \times 0.49^2 + 3 \times 0.51^2 \times 0.49 \\ &= 13.27\% + 36.74\% + 38.23\% + 11.76\%.\end{aligned}$$

2. **The Poisson distribution.** In binomial distribution we study the possibility of events occurred and not occurred. But there are certain events in which possibility of event occurred is only important. For example in a football match how many goals scored is important. Death due to heart attack in a year in a town may only be recorded.

Thus Poisson distribution is limiting case of Binomial distribution. When p is a very small number and n is so big that np is finite constant, means equal to m. According to this distribution probability of r successes is as follows :

$$p(X=r)=\frac{m^r e^{-m}}{r\,!}$$

Here m is Parameter of distribution

r = 0, 1, 2, 3 .... accepts the values

! = Factorial.

We know from Binomial expression that step —

$$f(r) = nc_r p^r q^{n-r}$$

$$= \frac{n(n-1)(n-2)\ldots(n-r+1)}{r\,!}\,.\,p^r q^{n-r}$$

$$= n.n\left(1-\frac{1}{n}\right)^n\left(1-\frac{2}{n}\right)\ldots n\left(1-\frac{r-1}{n}\right)\times p^r(1-p)^{n-r}$$

because $p+q=1 \quad \therefore \quad q=1-p$

$$= \frac{\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{r-1}{n}\right)}{r\,!}\times n^r p^r (1-p)^{n-r}$$

$$= \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{r-1}{n}\right)\times(np)^r\left(1-\frac{np}{n}\right)^{n-r}$$

Now taking limit when $n \to \infty$ and $p \to 0$ by which $np = m$, we find

$$f(r) = \frac{1}{r\,!}\times m^r \times \underset{n\to\infty}{\text{Lt}}\left(1-\frac{m}{n}\right)^{n-r} \qquad \ldots(i)$$

But $\quad \underset{n\to\infty}{\text{Lt}}\left(1-\frac{m}{n}\right)^{n-r} = \underset{n\to\infty}{\text{Lt}}\left(1-\frac{m}{n}\right)^n \times \underset{n\to\infty}{\text{Lt}}\left(1-\frac{m}{n}\right)^{-r}$

$$= e^{-m}\times 1 = e^{-m}$$

$\therefore$ From (1)

$$f(r) = \frac{1}{r\,!}\,m^r e^{-m}$$

i.e. probability of r success when $n \to \infty$ and $p \to 0$

$$np = m = \frac{1}{r\,!}\,m^r e^{-m}$$

It mean $\quad P(X=r) = \dfrac{1}{r\,!}\,m^r e^{-m}.$

Thus, $r = 0$, keeping 1, 2, 3, 4 values of 0, 1, 2, 3 probabilities is

$$e^{-m}, \frac{me^{-m}}{1\,!}, \frac{m^2e^{-m}}{2\,!}, \frac{m^3e^{-m}}{3\,!} \ldots$$

In Tabular form this can be kept as :

x, How many times	*Frequency*
0	e^{-m}
1	$\frac{me^{-m}}{1\,!}$
2	$\frac{m^3e^{-m}}{2\,!}$
3	$\frac{m^3e^{-m}}{3\,!}$
....	....
r	$\frac{m^re^{-m}}{r\,!}$
....	....

Example. Attribute one Poisson distribution and calculate theoretical frequency :

Death	0	1	2	3	4
Frequency	122	60	15	2	1

Calculation :

$$\text{(Mean)} \quad i.e. \quad m = \frac{122 \times 0 + 60 \times 1 + 15 \times 2 + 2 \times 3 + 1 \times 4}{122 + 60 + 15 + 2 + 1}$$

$$= \frac{0 + 60 + 30 + 6 + 4}{200} = \frac{100}{200} = \frac{1}{2} = 0.5$$

Now $$e^{-m} = \left\{1 - m + \frac{m^2}{2\,!} - \frac{m^3}{3\,!} + \ldots\right\}$$

$\therefore$ $$e^{-0.5} = 1 - (0.5) + \tfrac{1}{2}(0.5)^2 - \tfrac{1}{2}(0.5)^3 + \ldots$$
$$= 1 - 0.5 + 0.125 - 0.0208 + \ldots$$
$$= 0.61 \text{ approximate.}$$

$\therefore$ Probability of death (0) $= e^{-m} = e^{-0.5} = 0.61$

Probability of death (1) $= \frac{me^{-m}}{1\,!} (0.5)\,(0.61) = 0.305$

$$\text{Probability of death (2)} = \frac{m^2 e^{-m}}{2\,!} = \frac{(0.5)^2 (0.61)}{2} = 0.076$$

$$\text{Probability of death (3)} = \frac{m^3 e^{-m}}{3\,!} = \frac{(0.5)^3 (0.61)}{3} = 0.025$$

$$\text{Probability of death (4)} = \frac{m^4 e^{-m}}{4\,!} = \frac{(0.5)^4 (0.61)}{4} = 0.009.$$

$$\text{Frequency} = n e^{-m} = 200 \times 0.61 = 122$$

1. Death theoretical frequency

$$= \frac{N.m e^{-m}}{1\,!} = 200 \times .305 = 61$$

2. Death theoretical frequency

$$= \frac{N.m^2 e^{-m}}{2\,!} = 200 \times .076 = 15$$

3. Death theoretical frequency

$$= \frac{N.m^3 e^{-m}}{3\,!} = 200 \times 0.025 = 5$$

4. Death theoretical frequency

$$= \frac{N.m^4 e^{-m}}{4\,!} = 200 \times 0.009 = 1.8.$$

3. The normal distribution. Both Binomial and Poisson distributions are discontinuous distribution which are concerned to only complete numbers. For example number of persons in a family, number of death in a hospital, number of birth in a locality etc. But we come across a number of biological measurements where the variables are continuous in nature. They can be adequately described only by a continuous probability distribution.

"Normal distribution is a kind of mathematical distribution which deals with all continuous changing variables." This distribution is concerned to all numbers whether complete or in fraction. In normal distribution generally maximum cases fall in the middle of the series. For example if we study the height of Indian, the height of most of the Indian will fall between 150 cm – 180 cm. Height of very few Indian will fall between 90 cm – 120 cm (lower end of scale) and between 180 cm – 210 cm (higher end of scale).

The normal distribution curve. The curve made with the help of data of normal distribution is called normal distribution curve or Gaussian curve. The normal distribution curve is mostly bell shaped and therefore also called as bell shaped curve.

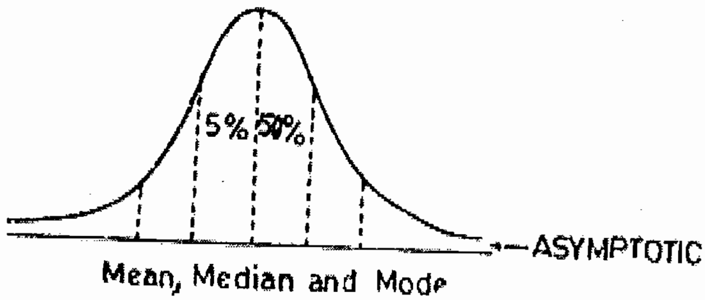


Fig. 8.3. Normal distribution curve or Gaussian curve.

Properties of normal distribution curve :

(1) The normal distribution curve is a continuous curve and is associated with continuous variables like weight, leng.h, age etc.

(2) The curve contains a peak and is symmetrical and asymptotic (touches at infinity).

(3) All measures of central tendency are equal and stable on the highest peak axis *i.e.*

Mean = Median = Mode.

(4) The normal distribution curve has a fixed mathematical characteristic feature independent of the scale (unit of measurement), magnitude and unit of mean and standard deviation.

(5) Maximum observations of distribution lie in the middle of curve. Few observations lie on left and right end of the curve.

(6) The middle area of normal curve and axis is known as area of curve and indicates the maximum number of frequency distribution. On the basis of mean and standard deviation, area of curve of normal distribution is as follows :

(*a*) Mean ± 1 S.D. – 68.268% Relative frequency.
(*b*) Mean ± 2 S.D. – 95.45% Relative frequency.
(*c*) Mean ± 3 S.D. – 99.73% Relative frequency.

Each side of mean has 49.865% relative frequency.

(*d*) Mean ± 3 S.D. covers almost entire area of curve. Only .27% area remains outside of curve.

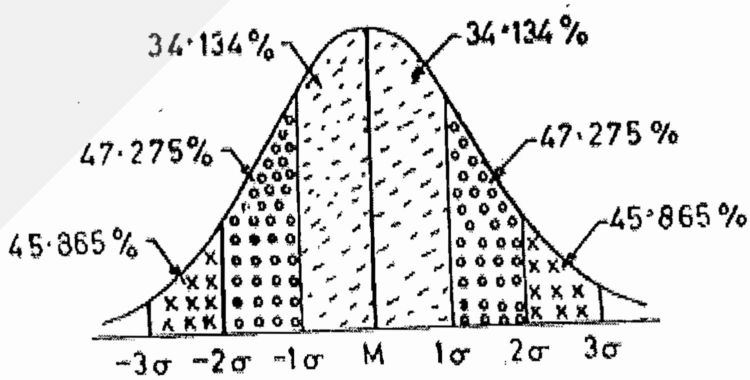


Fig. 8.4.

Fig. 8.4 shows probability or relative frequency of an observation falling beyond mean ± 2 S.Ds and ± 3 S.Ds respectively.

Besides this, following fact are also to be considered in context to curve area :

$\bar{x} \pm 1.96\sigma$ covers 95% area.
$\bar{x} \pm 2.578\sigma$ covers 99% area.
$\bar{x} \pm .6745\sigma$ covers 50% area.

(7) An area bounded by a distance of 0.67449 S.D. on each side of the mean (Fig. 8.4) will include exactly half of the observations. So for a single observation chances are exactly equal for deviation from the mean by an amount greater or less than 0.67449 S.D.

This fact is used to calculate probable error :

We can say that a value of X whose deviation from the mean either positively or negatively, is greater than 1 S.D., will occur 1 in 3 times.

A value with either +ve or −ve deviation greater than 2 S.D., will occur 1 in 20 times. So either +ve or −ve value of deviation greater than 3 S.D. will occur 1 in 370 times.

On these facts is based the concept of significance which indicate "that the odds are heavy against the deviation from its expected value of a particular estimate."

Odds of 19 to 1 against an occurrence by chance are usually taken as significance of that occurrence. "This coincides roughly to the odds of getting deviation from the mean of a normal distribution greater than twice the S.D. either +vely or −vely.

Test of the normal distribution. Various statistical calculations depend on normal distribution. Test for normal distribution can be done by simple mathematical calculations :

(*a*) To test the normality of the distribution in which frequency (f) and the values of distribution X are given.

(*b*) Prepare 4 more columns headed of fX, fX^2, fX^3 and fX^4.

(*c*) Add all the columns to give $\Sigma f = N$ and ΣfX, ΣfX^2, ΣfX^3 and ΣfX^4.

(*d*) Divide the sum of each last four columns by N to get the four quantities called V_1, V_2, V_3 and V_4.

(*e*) Obtain the four moments about the mean from the equations :

$$u_1 = 0$$
$$u_2 = V_2 - V_1^2$$
$$u_3 = V_3 - 3V_1.V_2 + 2V_1^3$$
$$u_4 = V_4^3 - 4V_1.V_3 + 6V_1^2.V_2 - 3V_1^4.$$

(*f*) When the variate is continuous certain correction have to be applied and the equations become :

$$u_1 = 0$$

$$u_2 = V_2 - V_1^2 - \frac{1}{12}$$

$$u_3 = V_3 - 3V_1.V_2 + 2V_1^3$$

$$u_4 = V_4^3 - 4V_1.V_3 + 6V_1^2.V_2 - 3V_1^4 - \frac{1}{2}u_2 - \frac{1}{80}$$

(*g*) Now calculate the B_1 and B_2 two constants from the formulae :

$$B_1 = \frac{u_3^2}{u_2^3},$$

$$B_2 = \frac{u_4}{u_2^2}.$$

(*h*) Now obtain the two quantities Y_1 and Y_2 which are related to B_1 to B_2 from the following formulae :

$$Y_1 = \pm\sqrt{B_1}$$

$$Y_2 = B_2 - 3.$$

Y_1 = is a measure, whether or not the distribution is symmetrical.

Y_2 = measures the departure of a symmetrical nature from normality.

(*i*) Now calculate the S.E. of Y_1 and Y_2 by $\sqrt{\frac{6}{N}}$ and $\sqrt{\frac{24}{N}}$ respectively.

(*j*) If the value of Y_1 and Y_2 are less than twice these standard errors, then the distribution is not significantly different from the normal form.

(*k*) If the value of Y_1 and Y_2 are greater than twice their standard error, the distribution is not normal.

Example. Test the normality of the following distribution showing '*f*' and '*x*'.

f	3	5	11	21	41	101	45	23	11	7	3
x	–6	–5	–4	–3	–2	0	2	3	4	5	6

Following Table 8.3 is prepared on the basis of above data :

Table 8.3.

f	x	fx	fx^2	fx^3	fx^4
3	−6	−18	108	−648	3888
5	−5	−25	125	−625	3125
11	−4	−44	176	−704	2816
21	−3	−63	189	−567	1701
41	−2	−82	164	−328	656
101	0	−232	0	−2872	0
45	2	90	180	360	720
23	3	69	207	621	1863
11	4	44	176	528	2112
7	5	35	175	875	4375
3	6	18	108	648	3888
$\Sigma f =$ 271		+256 −232	1688	+3032 −2872	25144
		+24		160	

Now
$$V_1 = \frac{24}{271} = 0.088$$

$$V_2 = \frac{1688}{271} = 6.228$$

$$V_3 = \frac{160}{271} = 0.590$$

$$V_4 = \frac{25144}{271} = 92.78$$

So,
$$\begin{aligned} U_2 &= 6.228 - .088^2 - 0.0833 \\ &= 6.228 - 0.00774 - 0.0833 \\ &= 6.228 - 0.0910 \\ &= 6.137 \end{aligned}$$

$$\begin{aligned} U_3 &= 0.590 - (3 \times .088 \times 6.228) + 2 \times 0.088^3 \\ &= 0.590 \times 1.644 + 2 \times 0.000681 \\ &= -1.054 + 0.00136 \\ &= -1.05264. \end{aligned}$$

$$\begin{aligned} U_4 &= 92.78^3 - 4 \times 0.088 \times 0.590 + 6 \times 0.088^2 \times 6.228 - 3 \times 0.88^4 \\ &= 90.6573 - 3.0806 = 87.5767. \end{aligned}$$

Now calculate B_1 and B_2

$$B_1 = \frac{u_3^2}{u_2^3} = \frac{-(1.05264)^2}{(6.137)^3} = \frac{1.1080}{231.1364} = 0047937$$

$$B_2 = \frac{u_4}{u_2^2} = \frac{87.5767}{6.137^2} = \frac{87.5767}{37.6627} = 2.5325$$

So $$Y_1 = \pm\sqrt{B_1} = \sqrt{.0047937} = .0692365$$

$$\text{S.E.} = \sqrt{\frac{6}{N}} = \sqrt{\frac{6}{271}} = 0.1487$$

$$Y_2 = B_2 - 3 = 2.532 - 3 = -.461$$

$$\text{S.E.} = \sqrt{\frac{24}{N}} = \sqrt{\frac{24}{271}}$$

$$= \sqrt{.08856} = .2975.$$

Inference. As $Y_1 = .0692$ is less than twice its S.E. $(2 \times .148)$, so the distribution is symmetrical.

$Y_2 = -.461$ is slightly less than twice its SE. (2×0.29). Therefore the distribution is almost symmetrical.

Measures of deviations from the normal distribution. *Skewness and curtosis*—We have studied that most of the cases fall in the middle of normal distribution curve. But often normal distribution curve becomes different. Divergence from normal distribution curve (Bell shaped curve) may be of two types :

(1) **Skewness :** In skewed curve, mean, median and mode do not fall in the middle of the normal distribution curve. They fall on different points and centralization of scores fall on either left or right side. Mean, median and Mo is the same in normal distribution and skewness is zero. But in skewed distribution mean and median fall on different points. Skewness is of two types :

(*i*) Positive skewness and (*ii*) Negative skewness.

(*i*) **Positive skewness** is found in that type of distribution where centralization of scores remain on the left end of the scale.

(*ii*) **Negative skewness** is found in that type of distribution where centralization of scores remain on the right end of the scale.

(2) **Kurtosis :** If frequency distribution is more peaked or flat in comparison to normal curve, then the curve is called kurtosis. On graph peaked curve is called *leptokurtic* and flat curve is called *platykurtic*.

Though normal distribution curve is bell shaped, but the real form is determined by its mean and standard deviation. Following figures may help to understand the said statement.

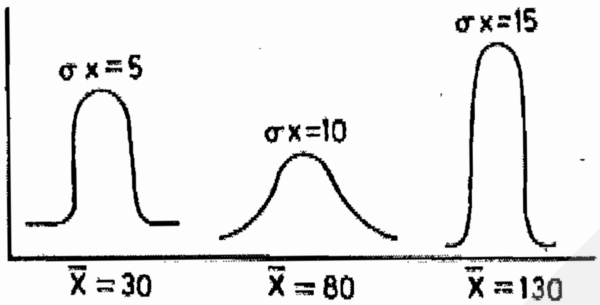


Fig. 8.5. Normal distribution curves with different mean and standard deviation.

Equation for normal distribution curve is as follows :

$$\gamma = \frac{1}{\sqrt{2\pi\sigma x^2}} \cdot e^{-1/2} \frac{(x-M)^2}{\sigma x}$$

Integration $$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma x^2}} e^{-1/2} \frac{(x-M)^2}{\sigma x} dx = 1.$$

Here γ = height of curve on X axis.

e = symbol for base of Neprian antilogrithm. Its value is 2.7183.

π = circumference of a circle in ratio of diameter. Its value is 3.1416.

∞ = Symbol of infinity.

M = Mean.

In the above equation if Mean and Standard deviation are known then normal curve can be obtained because other coefficients are constant.

Standard normal distribution is probability distribution of variable z, which keeps normal distribution with zero (0) Mean and unit variance distribution.

$$z \sim (0.1)$$

Gauss tried to exhibit the probabilities of different values of z through normal distribution table.

Standardisation procedure can be shown as follows :

If a variable x is normal distribution and their mean (μ) and variance σx^2 are known, then

$$z_i = \frac{x_i - M}{\sigma x} \sim N(0, 1)$$

Here x_i = That value fo x whose distribution has to be converted into Standard normal curve.

M = Mean of x distribution
σx = Standard deviation of x distribution.

EXERCISE

1. Two women gave birth to children simultaneously. Find out probability of issue of female to both women.
2. One event has to happen out of two. If probability of first is 2/3 of 2nd, then find out probability of first.
3. If A and B are two events, then calculate P(B/A).
   If (*a*) A is subset of B.
   (*b*) A and B are mutually exclusive.
4. What is the difference between observed frequency distribution and theoretical frequency distribution. Describe different types of theoretical distribution.
5. What do you understand by Binomial distribution ? 5 notes of one rupee were dropped 1000 times and each time noted the number of head which is as follows :

X	0	1	2	3	4	5	
(f)	38	144	342	287	164	25	= 1000

   Test whether distribution is binomial. If it is so find mean and Standard Deviation.
6. What do you mean by Poisson distribution. In one Poisson distribution P(x) = .1 when x = 1, then find out mean.
7. Fit in Poisson distribution in following distribution :

Incorrect number	0	1	2	3	4	5	6	7	8	9	10
Page number in book		4	15	22	21	20	8	6	2	0	1

8. What do you mean by normal distribution ? Suppose height of plants are in normal distribution. 95% are between 61″ and 74″, then find out mean and standard deviation of distribution of height of plants.
9. In a normal distribution of things 31% are below 45 and 8% above 64. Find out mean and standard deviation of distribution.
10. Two groups of sheeps denoted as A and B contain two white and one black sheeps, one white and five black sheeps. One sheep is transferred from A to B and then one sheep is drawn at random from the latter. It happens to be white. What is the probability that the transferred sheep was black ?

9. Correlation

Synopsis. *Introduction ; Types of correlation ; Correlation coefficient ; Methods of studying correlation—Scatter diagram method. Pearson's product moment method and Spearman's Coefficient of Rank correlation ; Standard error of the correlation coefficient and verification of significance of correlation coefficient. Exercise.*

Introduction. Literally correlation means association of two or more facts. In statistics correlation may be defined as '*the tendency of simultaneous variation between two variables.*' The distribution involving two variables are called *bivariate distribution* and the distribution involving more than two variables are called *multivariate distribution.* In statistics we study the degree of correlation between two or more variables. Sometimes two variables are measured in the same individual such as length and weight, oxygen consumption and body weight. Body weight and Hb% etc. At other times the same variable is measured in two or more related groups such as tallness in parents and offspring, intelligence quotient (IQ) in brothers and in corresponding sisters (siblings) and so on.

In a bivariate distribution, the correlation may be positive or negative, and linear or curvilinear.

Two variables co-varying in the same direction are positively correlated. For example, we expect a positive correlation between height and weight of a group of individuals.

Covariation between the two variables in opposite direction are negatively correlated. The increase in one variable results in a corresponding decrease in the other. For example, increase in number of caterpillar results in a corresponding decrease in number of leaves of plants.

The correlation of two variables which can be expressed by a straight line is called linear correlation. In perfect linear correlation the amount of change in one variable bears a constant ratio to the amount of change in the other. For example length of 5 fishes of a species and their snout length is measured in cm. The measurement is given below :

Body length : 8, 9, 11, 12, 13 X variable
Snout length : 1, 2, 4, 5, 6 Y variable

The above observation indicates that each individual score 1 cm more on test Y. This mean that the correlation between the above two variables

is expressible in the form Y = X + 1, which is an expression representing a straight line, *i.e.* a perfect positive linear relationship, in which correlation between X and Y will be +1. (Though it rarely happens in biological experiments).

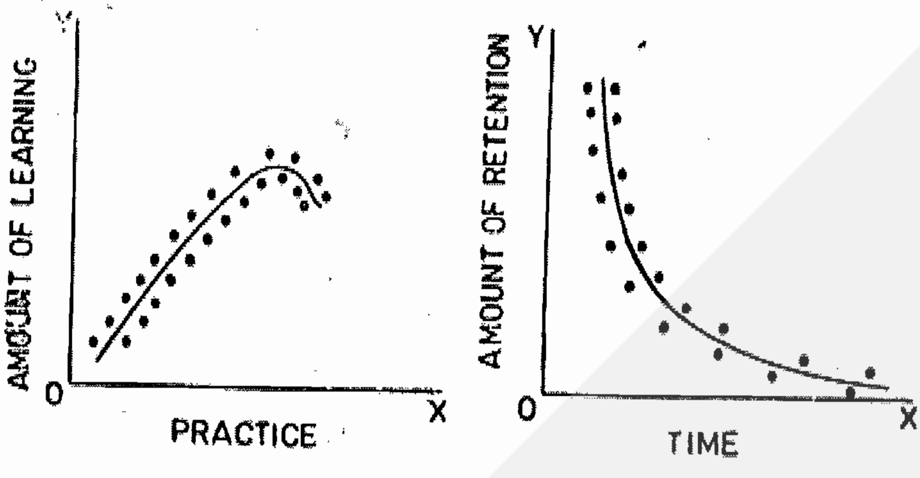


Fig. 9.1. (A) Correlation between practice and amount of learning. (B) Correlation between time and amount of retention.

The correlation of cores of two variables based on some quality shown by a curve line on graph is called curvilinear correlation. Above Fig. 9.1 (*a*) shows the correlation between practice and amount of learning and (*b*) shows the correlation between time and amount of retention. Graph (*a*) reveals that upto a certain limit amount of learning increases with the practice, but after a certain limit capacity of learning fall even if practice remains same or continued. Graph (*b*) indicates that correlation between time and amount of retention is represented by curve line. It indicates that with the lapse of time power of retention (memory) decreases upto a certain limit. After that memory even faint remains present although time passes.

Correlation coefficient. Numerical expression of correlation is called mathematical correlation. In other words, correlation of two variables by mathematical method is obtained by correlation coefficient. In biological experiments use of correlation coefficient is very significant. According to J.P. Guilford "*A coefficient of correlation is a single number that tells us to what extent two or more things are related and to what extent variations in one go with variations in other.*" The correlation coefficient is expressed by a letter of English '*r*'.

Methods of studying correlation :

There are three methods of studying correlation between two variables in the case of ungrouped data :

1. Scatter diagram method.
2. Pearson's product moment method and
3. Spearman's coefficient of Rank correlation.

1. Scatter diagram method. Scatter diagram or dot diagram is a graphic device for drawing certain conclusions about the correlation between two variables. In preparing a Scatter diagram, the observed pairs of observations are plotted by dots on a graph paper by taking the measurements on variable X along the horizontal axis and that on variable Y along the vertical axis. The placement of these dots on the graph reveals the change in the variable as to whether they change in the same or in opposite directions. Scatter diagram showing various degrees of correlation and it is shown in Fig. 9.2 (*a*), (*b*), (*c*), (*d*), (*e*).

[**Note.** This book is confined to bivariate distribution and ungrouped data in context to correlation.]

Types of correlation. There may exist five kinds of correlation between two variables depending on its extent and direction. Each type may be shown both mathematically and graphically :

(*i*) **Perfect positive correlation.** The two variables denoted by letter X (Body length) and Y (Body weight) are directly proportional and fully correlated with each other. Both variables rise or fall in the same proportion. Examples of perfect or total correlation is very very rare in nature but some approaching to that extent are there such as *day length* and *temperature* ; *rain and humidity* ; *body weight* and *height* ; *age* and *height* ; *age* and *weight* etc. upto certain age. The imaginary mean line rising from the lower ends of both X and Y axes forms a straight line. When scatter diagram is drawn all the points fall around the mean line [Fig. 9.2 (*a*)].

(*ii*) **Moderately positive correlation.** The two variables denoted by X (Age of husband) and Y (Age of wife) are partially positively correlated. Values of correlation coefficient (r) lie between 0 and +1, *i.e.*, $0 < r < 1$. Other examples of positive correlation may be infant mortality rate and overcrowding, tallness of plants and the quantity of manure used, nutrition and death rate in pregnancy etc.

In such moderately positive correlation, the scatter will be there around an imaginary mean line, rising from the lower ends of both X and Y variables [Fig. 9.2 (*c*)].

(*iii*) **Perfect negative correlation.** The variables denoted by letter X (Temperature) and Y (Lipid content of body of a species of fish) are inversely proportional to each other, *i.e.*, when one (X) rises the other (Y) falls in the same proportion. The correlation coefficient (r) = −1 to 0. Examples of perfect negative correlation is also very rare in nature but some approaching to that extent are there such as temperature and lipid content of the body, RBCs number and Hb%, T_4 injection and oxygen

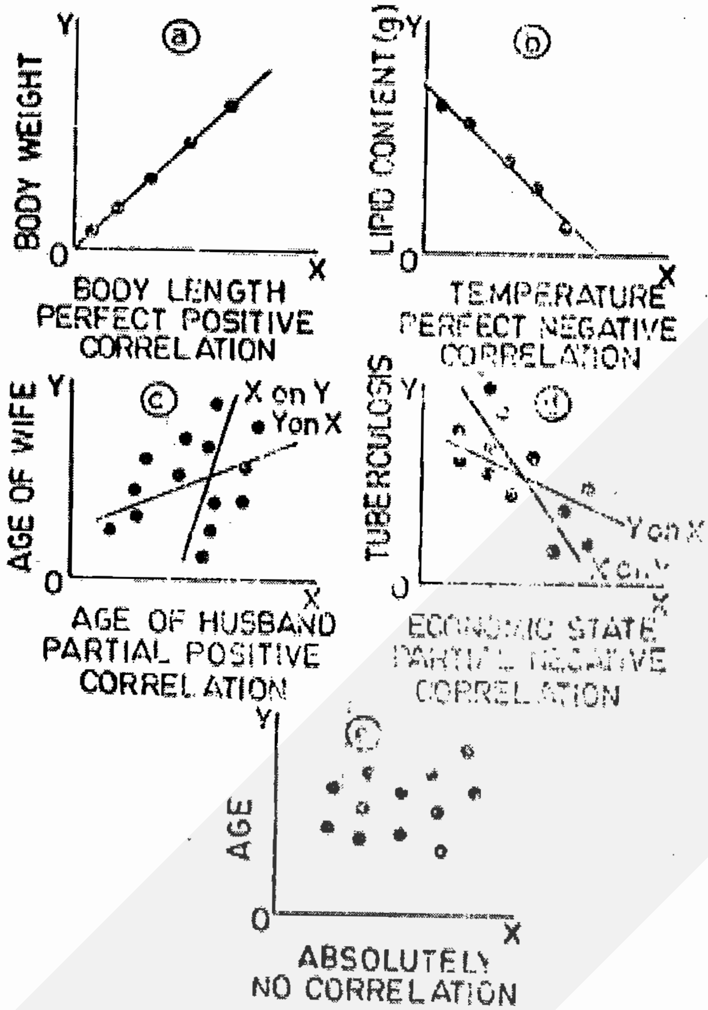


Fig. 9.2. Different types of correlation.
(*a*) Perfect positive correlation ; (*b*) Perfect negative correlation ; (*c*) Partial positive correlation ; (*d*) Partial negative correlation ; (*e*) Absolutely no correlation.

(*iv*) **Moderately negative correlation.** The two variables denoted by X (Economic condition of State) and Y (case of Tuberculosis). In this case values of correlation coefficient lie between -1 and 0 such as income and infant mortality rate, age and vitality in adults etc.

In such moderately negative correlation, the scatter will be there around an imaginary mean line rising from the extreme values of variable [Fig. 9.2 (*d*)].

(*v*) **Absolutely no correlation.** In this case the value of correlation coefficient (r) is zero, indicating that no linear relationship exists between the two variables. There is no imaginary mean line indicating trend of correlation. X is completely independent of Y such as Hb% and body weight,; Body weight and IQ etc.

In absolutely no correlation X variable is completely independent of Y variable. In this case points are so scattered that no imaginary line can be drawn. [Fig. 9.2 (*e*)].

2. Pearson's product moment method. It is also known as Pearson's coefficient of correlation. It is one of the most widely used algebraic methods of finding correlation between two variables. The coefficient of correlation (r) gives an idea about the degree of linear relationship between two variables. Formula to obtain coefficient of correlation (r) is used as follows :

$$r = \frac{\Sigma x.y}{\sqrt{\Sigma x^2.y^2}}$$

where X is the independent variable normally represented by the abscissa and Y is the dependent variable represented by the ordinate. x and y are the deviations from the respective means (as used for other purposes—determination of variance and standard deviation).

In language we can say that 'r' can be calculated by dividing the sum of products of deviations from their respective means by the square root of the products of the sums of squares of deviations from the respective means of the two variables.

Here r = correlation coefficient
x = deviations of X variable
y = deviations of Y variable
$\Sigma x.y$ = sum of multiplication of deviations x and y.

Pearson's product moment method is applied is 4 ways :

(*i*) **With the help of actual mean.** The above formula is applied when 'r' is obtained applying actual mean.

Example. The length and weight of 7 groups of fishes of a species is given below. Find out correlation coefficient of the two variables.

Length of body. 11.7 cm, 13.9 cm, 15.5 cm, 17.8 cm, 18.5 cm, 19.2 cm, 21 cm

Weight of body. 7.10 g, 12.42 g, 15.35 g, 23.20 g, 28.45 g, 32.25 g and 39.84 g.

Table 9.1.

S. No.	*Length* X	*Weight* Y	x	y	x^2	y^2	$x.y$
1.	11.7	7.10	−5.1	−15.58	26.01	241.8	79.3
2.	13.9	12.42	−2.9	−10.236	8.41	104.6	23.36
3.	15.5	15.35	−1.3	−7.33	1.69	53.2	9.49
4.	17.8	23.20	+1	+.55	1.0	.30	.55
5.	18.5	28.45	+1.7	+5.8	2.89	33.64	9.86
6.	19.2	32.45	+2.4	9.6	5.76	92.16	28.8
7.	21	39.84	+4.2	+17.19	17.64	295.49	72.2
N = 7	ΣX = 117.6	ΣY = 158.81			Σx^2 = 66.64	Σy^2 = 821.19	$\Sigma x.y$ = 223.56

Calculation. For calculation of correlation coefficient from above data (ungrouped series) a table is prepared with the help of following steps :

— Make a table of 8 columns.
— Mention serial numbers in column 1, value of X in column 2 and value of Y in column 3.
— Find out Actual mean of X and Y with the help of formula $-\Sigma X/N$.
— find out deviation of all scores of X and Y. Formula to find out deviation from actual mean is $X - \overline{X}$. (Score—mean). Put all values of deviations in column 4 against their scores for variable X and in column 5 for variable Y.
— Find the square of x and y and put them in column 6 and 7 respectively.

Put the multiplication value of x and y of each score in last column *i.e.* 8th column. All values of $x.y$ is summed and given in last column of Table 9.1.

$$\overline{X}_1 \text{ of } X = \frac{117.6}{7} = 16.8. \quad \text{Mean of length of fish}$$

$$\overline{X}_2 \text{ of } Y = \frac{158.81}{7} = 22.68. \quad \text{Mean of weight of fish}$$

$$r = \frac{\Sigma x.y}{\sqrt{\Sigma x^2.y^2}}$$

$$= \frac{223.56}{\sqrt{66.64 \times 821.19}}$$

$$= \frac{223.56}{\sqrt{54724.101}}$$

$$= \frac{223.56}{233.93} = 0.96.$$

Inference. The calculated value of correlation coefficient (r) is 0.96. One has to see the significance of 'r' at .05 and 0.01 level. First of all we find out df. Here df = N − 2 *i.e.* 7 − 2 = 5.

Here df = 5 and calculated value of r = .96.

On verification of correlation table we observe that value of 'r' at *d.f.* 5 is .755 at .05 level. Calculated value of 'r' is 0.96. Since the calculated value is higher, therefore it is clear that 'r' is significant at .05 level at df 5. Now we can safely say that both variables *i.e.* length and weight of body is in complete +ve correlation.

(*ii*) **With the help of assumed mean.** Correlation coefficient 'r' is obtained with the help of following formula where assumed mean is used :

$$r = \frac{\frac{\Sigma x'.y'}{N} - CX.CY}{\sigma x'.\sigma y'}$$

Here r = correlation coefficient
x' = deviations obtained from assumed mean in X variable
y' = deviations obtained from assumed mean in Y variable
$\sigma x'$ = standard deviation multiplied by x' of X variable
$\sigma y'$ = standard deviation multiplied by y' of Y variable
cX = correction of X variable
cY = correction of Y variable.

Example. Find out 'r' with the help of above formula (assumed mean method).

To obtain the values required for formula a Table 9.2 is prepared using data of previous example.

Table 9.2.

S. No.	*Length X*	*Weight Y*	x	y	x^2	y^2	$x.y$
1.	11.7	7.10	−6.1	−16.1	27.21	259.21	98.21
2.	13.9	12.42	−3.9	−10.78	15.25	116.20	42.42
3.	15.5	15.35	−2.3	−7.85	5.29	61.62	18.05
4.	17.8	23.20	0	0	0	0	0
5.	18.5	28.45	+.6	+5.25	36	27.56	3.15
6.	19.2	32.45	+2	9.05	4	81.90	18.1
7.	21.0	39.84	+3.2	+16.64	10.24	276.88	53.24
N = 7	ΣX = 117.6	ΣY = 158.6			Σx^2 = 107.95	Σy^2 = 823.37	$\Sigma x.y$ = 233.17

$$\bar{X} = \frac{117.6}{7} = 16.8$$

$$\bar{Y} = \frac{158.6}{7} = 22.65$$

Here we have assumed that score of 4th serial of X and Y variable are their means.

$$\bar{X} = 17.8$$
$$\bar{Y} = 23.20.$$

Assumed mean for X = 17.8 which falls in the middle of X variable.

Assumed mean for Y = 23.20 which falls in the middle of Y variable column.

CX = Actual mean of X − Assumed mean of X.

$$= 16.8 - 17.8 = -1$$
$$C^2X = (-1)^2 = 1$$

$$CY = \text{Actual mean of Y} - \text{Assume mean of Y}$$
$$= 22.65 - 23.20 = 0.55$$
$$C^2Y = (0.35)^2 = .3025$$

$$\sigma x' = \sqrt{\frac{\Sigma x'^2}{N} - C^2X}$$

$$= \sqrt{\frac{107.95 - 1}{7}}$$

$$= \sqrt{15.42 - 1} = \sqrt{14.42}$$

$$= 3.797$$

$$\sigma y' = \sqrt{\frac{\Sigma y'^2}{N} - C^2Y}$$

$$= \sqrt{\frac{823.37}{7} - .3025}$$

$$= \sqrt{117.62 - .3025}$$

$$= \sqrt{117.31} = 10.83.$$

Fit in the values calculated in following formula :

$$r = \frac{\frac{\Sigma x'.y'}{N} - CX.CY}{\sigma x'.\sigma y'}$$

$$= \frac{\frac{233.17}{7} - (-1).(.55)}{2.797 \times 10.83}$$

$$= \frac{33.31 + .55}{41.12} = 0.82.$$

Inference. To test the significance of 'r' at $d.f.$ = N – 2 = 7 – 2 = 5. On .05 or .01 level.

Table 'r' reveals that value of r at 5 $d.f.$ on .05 level is .755. The calculated value of 'r' is 0.82 which is above than the tabulated value.

Therefore we can say that the length and weight of the fish has high correlation in positive direction.

(*iii*) **Using raw score.** Above two methods to obtain 'r' is time taking. We can find out 'r' with the help of following formula using raw score without finding deviation.

$$r = \frac{N/\Sigma X.Y - \Sigma X.\Sigma Y}{\sqrt{N\,[\Sigma X^2 - (\Sigma X)^2] \times [N\Sigma Y^2 - (\Sigma Y)^2]}}$$

Example 1. Body weight (g) and body water per cent out of 12 fishes of same species are recorded as follows. Obtain 'r' of above data :

Body weight in (g)	*Percentage of water*
27.2	62.6
31.5	64.2
32.6	67.0
29.2	68.6
28.3	71.0
30.3	72.6
29.9	71.0
31.8	69.7
32.4	67.0
28.6	64.5
31.8	62.0
27.1	60.7

A table 9.3 is prepared from the above data.

Table 9.3

Body wt. *X*	*% of water* *Y*	X^2	Y^2	$\Sigma X.Y$
27.2	62.6	739.84	3918.76	1702.72
31.5	64.2	992.25	4121.64	2022.3
32.6	67.0	1062.76	4489.00	2184.2
29.2	68.6	852.64	4705.96	2003.12
28.3	71.0	800.89	5041.00	2009.3
30.3	72.6	918.09	5270.76	2199.78
27.9	71.0	778.41	5041.00	1980.9
31.8	69.7	1011.24	4858.09	2216.46
32.4	67.0	1049.76	4489.00	2170.8
28.6	64.5	817.96	4160.25	1844.7
31.8	62.0	1011.24	3844.00	1971.6
27.1	60.7	734.41	3684.49	1644.97
$\Sigma X = 358.7$	$\Sigma Y = 800.9$	$\Sigma X^2 =$ 10769.49	$\Sigma Y^2 =$ 53623.95	$\Sigma X.Y =$ 23950.85

$$r = \frac{N\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{(N)\,[\Sigma X^2 - (\Sigma X)^2] \times [(N)\,\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{12 \times 23950.85 - (358.7) \times (800.9)}{\sqrt{(12)\,(10769.49) - (358.7)^2] \times [12 \times 53623.95 - (800.9)^2]}}$$

$$= \frac{287410.2 - 287282.83}{\sqrt{[129233.88 - 128665.69]\ [643487.4 - 641440.81]}}$$

$$= \frac{127.37}{\sqrt{(568.19)\ (2046.59)}} = \frac{127.37}{\sqrt{1162851.9}} = \frac{127.37}{1078.35} = 0.118.$$

Here calculated value of $r = 0.118$.

$$df = 12 - 2 = 10$$
$$df_{10} = .576 \text{ at } 0.05 \text{ level}$$
$$df_{10} = .708 \text{ at } .01 \text{ level. (Appendix 5)}$$

Inference. The calculated value is very low than table value. The value of r suggests that the weight of fish and water % in body is not correlated. Though in nature which are correlated. This might have happened due to hypothetical data or experimental error.

[**Note.** This method is easy but can't calculated without calculator if value of X and Y exceeds 3 or 4 digits.]

(*iv*) **Applying difference method.** 'r' is obtained using actual mean and difference of deviations of both variable. Following formula is used for this purpose :

$$r = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma d^2}{2\sqrt{\Sigma x^2\ \Sigma y^2}}$$

Here, x = Deviation obtained from actual mean for X variable
y = Deviation obtained from actual mean for Y variable
d = Difference of x and y.

Example. Using the same data of raw score method (body weight and body water percentage of 12 fishes) a table 9.4 of 9 columns is prepared which can satisfy the above formula to deduce 'r'.

Table 9.4

S. No.	X	Y	x	y	$x - y$	d^2	x^2	y^2
1.	27.2	62.6	−2.69	−4.14	+1.45	2.10	7.23	17.13
2.	31.5	64.2	1.61	−2.54	+4.15	17.22	2.59	6.45
3.	32.6	67.0	2.71	0.26	+2.45	6.00	7.34	0.06
4.	29.2	68.2	−0.69	1.46	−2.15	0.32	0.47	2.13
5.	28.3	71.0	−1.59	4.26	−5.85	34.22	2.52	18.14
6.	30.3	72.6	0.41	5.86	−5.45	29.70	0.16	34.33
7.	29.9	71.0	0.01	4.26	−4.25	18.06	0.0001	18.14
8.	31.8	69.7	1.91	2.96	−1.05	1.10	3.64	8.76
9.	32.4	67.0	2.51	0.26	+2.25	5.06	6.300	0.06
10.	28.6	64.5	−1.29	−2.24	+0.95	0.90	1.66	5.01
11.	31.8	62.0	1.91	4.74	+6.65	44.22	3.64	22.46
12.	27.1	60.7	−2.79	−6.04	+3.25	10.56	7.78	36.48
	$\Sigma X =$ 358.7	$\Sigma Y =$ 800.9				$\Sigma d^2 =$ 169.46	$\Sigma x^2 =$ 43.33	$\Sigma y^2 =$ 169.15

360.7 800.5

$$r = \frac{\Sigma x^2 + \Sigma y^2 - \Sigma d^2}{2\sqrt{\Sigma x^2 . \Sigma y^2}}$$

$$= \frac{43.33 + 169.15 - 169.46}{2\sqrt{(43.33) - (169.15)}}$$

$$= \frac{212.48 - 169.46}{2\sqrt{7329.2695}}$$

$$= \frac{43.02}{2 \times 85.61111}$$

$$= \frac{43.02}{171.2222} = 0.2515. \quad \textbf{Ans.}$$

Inference. Find out significance of 'r' at $df = 12 - 2 = 10$.

df_{10} = 0.576 at 0.05 level.
df_{10} = 0.708 at 0.01 level.

Here calculated $r = 0.251$.

Calculated value is much less than table value. Therefore, value of r is less do not have at 0.05 level. Therefore we can say that body wt. and percentage of water of body do not have correlation.

3. Spearman's ran difference method. When two variables are correlated but they do not follow normal distribution then 'r' is calculated by Spearman's rank difference method. Its symbol is ρ (rho). In rank difference method N should be small (should not exceed 30). Following formula is used to calculate ρ (rho) :

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}.$$

Here, ρ (rho) is the rank difference of X and Y variables,

D = Difference of two corresponding observations in two variable.

ΣD^2 = Summation of square of difference of two variables rank I and II (R_1 and R_2).

Example. Number of fishes (X) and no. of helminth parasites (Y) were as follows. Find the rank correlation (Rho ρ).

X	17	17	18	19	19	20	21	22	23
Y	230	210	290	230	330	320	360	340	320

Following steps have to be taken to find ρ (Rho) :

(*i*) **To ascertain rank.** In X scores maximum *i.e.* 23, therefore, rank of this score is first (No. 1). Less score than this is 22 therefore rank of this score is second (No. 2). Likewise rank of score 21 is third (No. 3) and

20 is 4th. Score 19 is two times whose rank no is 5th and 6th. Their rank will be $\frac{5+6}{2} = 5.5$. Rank of score 18 will be 7th. Score 17 is two times therefore its rank no. is 8 and 9. Their rank will be $\frac{8+9}{2} = 8.5$.

[Suppose any score would have come thrice. Suppose score 17 have come three time whose rank no. is 8, 9 and 10 serially. Therefore rank no. for each 17 score will be $\frac{8+9+10}{3} = 9$.]

(*ii*) Rank of X observation is R_1 and Rank of Y observation is R_2.

(*iii*) Difference of R_1 and R_2 is calculated and the values are put in a column (D).

(*iv*) Square of R_1 and R_2 difference *i.e.* D^2 is calculated.

(*v*) All values of D^2 is summed up *i.e.* ΣD^2.

Following table 9.5 of six columns is prepared using data of above example.

Table 9.5.

X	*Rank*	*Y*	*Rank*	*D*	D^2
17	$\frac{8+9}{2} = 8.5$	230	$\frac{7+8}{2} = 7.5$	1	1
17	$\frac{8+9}{2} = 8.5$	210	9	–0.5	0.25
18	7	290	6	1	1
19	$\frac{5+6}{2} = 5.5$	230	$\frac{7+8}{2} = 7.5$	–2	4
19	$\frac{5+6}{2} = 5.5$	330	3	2.5	6.25
20	4	320	$\frac{4+5}{2} = 4.5$	–0.5	0.25
21	3	360	1	2	4
22	2	340	2	0	0
23	1	320	$\frac{4+5}{2} = 4.5$	–3.5	12.25
N = 9		N = 9			$\Sigma D^2 = 29$

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2-1)} = 1 - \frac{(6)\times(29)}{9(81-1)}$$

$$= 1 - \frac{174}{720} = \frac{720-174}{720} = \frac{546}{720} = .758. \quad \textbf{Ans.}$$

WEB: www.evolution.co.in | MAIL: info@evolution.co.in | PHONE: +91 11 4709 2329 ; + 91 9810 975 937 | ©Evolution Educare Pvt. Ltd.

Significance of ρ (Rho) is ascertained on verification of Spearman's rank difference appendix.

Here N = 9 and calculated value of ρ = .76.

Significance at 0.01 level on (9 – 1) = 8 d.f. is given as 0.712 in appendix. It means that fish and parasites has got positive correlation.

Standard error of the correlation coefficient and verification of significance of correlation coefficient.

'*r*' gives a measure of the degree of relationship between the two variables of a sample yet does not indicate the significance at the population level. There is also no possibility of deriving the significance value for the population directly. In order to test the deviation of '*r*' from zero we may apply the *t*-test by evaluating the ratio of '*r*' to the standard error of '*r*' with $n - 2$ degree of freedom. Standard error of '*r*' can be computed from *r* and *n* of the sample data.

$$\text{SE of } r = \sqrt{\frac{1-r^2}{n-2}}$$

Thus $$t = \frac{r}{\text{SE of } r} = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

or $$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$ for $n - 2$ degree of freedom.

Example. The value of '*r*' obtained in an example of Pearson's difference method in previous page where $r = .2515$.

$$t = \frac{0.2515\sqrt{12-2}}{\sqrt{1-(0.2515)^2}}$$

$$= \frac{0.2515\sqrt{10}}{1-0.0632}$$

$$= \frac{0.7952}{0.9368} = 848.$$

For $n - 2$ *i.e.* 10 *d.f.* at 5% level the highest value (tabulated value of *t*) is 1.812. The estimated value is .848 which is not much below the table value. Therefore the correlation between body weight of fish and water % of body is positively correlated.

10. Regression

Synopsis. *Introduction, Difference between correlation and Regression, Objects of Regression analysis, Linear Regression, Regression equation, Regression coefficient, Few examples, Calculation of Regression equation from values of deviation of mean of X and Y variables. Few examples of regression equation, Standard deviation for the regression line. Exercise.*

Introduction. F. Galton coined the term Regression in 1885 to explain the data obtained during the study of inheritance. Galton observed the height of offsprings of few generations of a family and came to the conclusion that the height of offsprings tend to remain in middle position. "The tendency to remain towards central position was called Regression by Galton."

We have studied that in order to draw a relationship, observations of two variables are plotted in the form of dots in a scatter diagram. A straight line is drawn which will approach as close as possible to all these points in the graph. The statistical analysis employed to find out the exact position of the straight line is known as the *linear regression analysis.* The main objective of regression analysis is to predict the value of one variable using the known value of the other. The existence of relationship between the independent variable X and the dependent variable Y can be expressed in a mathematical form known as the regression equation. The equation expressed by a straight line is called the linear regression equation.

Difference between Correlation and Regression :

(1) Correlation analysis tests the closeness and direction of relationship between the two phenomenon, whereas the regression analysis measures the nature and extent of this relation, thus enabling us to make prediction. Correlation coefficient is the measure of covariability between two variables while *regression* indicates the resultant relationship between independent and dependent variables so that prediction can be made of dependent variable for any value of independent variable.

(2) Correlation indicates the direction and quantity between two variables but it do not indicate that one variable is the cause and the other is result. Regression analysis indicates clearly the reason of relationship between two variables.

Objects of Regression analysis :

(1) The first aim or Regression analysis is to predict the value of one character of variable (variable say Y) from the known value of the other character or variable (variable say X). The former variable Y to be predicted is called dependent variable and the latter known variable X is called the independent variable. This is done by Regression line and by finding another constant called regression coefficient. Regression line explains the mean relationship between X and Y variables.

(2) To find out the measures of error, present during the use of regression line for prediction, is the another aim of Regression analysis. For this standard error of estimate is calculated.

Linear Regression. Linear relation between values of two variables are possible only when one unit change in the independent variable (X) influences change in definite quantity in dependent variables (Y). This change may be on the positive or negative side beyond the mean.

The lines of the best fit passing through the middle of points on plotted graph is drawn. These lines are called regression lines. Fig. 9.2 (*c*) and (*d*) on chapter 9. The two regression lines are drawn. one is X, Y and the other is Y on X indicating conditions of moderately +ve and moderately –ve correlation respectively. The two regression lines interest at the point where perpendiculars drawn from the means of X and Y variables meet.

When there is perfect correlation ($r = +1$ or -1) the two regression lines will coincide or become one straight line as in Fig. 9.2 (*a*) and (*b*) Chapter 9. Though perfect correlation is not possible in biological experiments. When the correlation is partial, the lines will be separate and diverge forming an acute angle at the meeting point of perpendiculars drawn from the means of two variables. Lesser the correlation, greater will be the divergence of angle. Steepness of the lines indicates the extent of correlation. Close the correlation greater is the steepness of regression lines X on Y and Y on X.

Composition of regression lines is based on least square assumptions. The general condition for regression lines is based on least square assumptions. The general condition for regression analysis is based from lines of the best fit is least. It is called least squared error.

Regression equation. The existence of relationship between the independent variable X and the dependent variable Y can be expressed in a mathematical form known as the **regression equation**. These equation represent the regression lines.

Regression equation of Y on X indicates the changes in the values of X for changes given in Y. Likewise regression equation of X and Y indicates the changes in the values of Y for changes given in X.

Regression equation of X on Y :

$$X = a + b.y$$

Regression equation of Y on X :

$$Y = a + b.x.$$

In both equations x and y are values of variables whereas a and b are constant. Constant a is intercepts *i.e.* it is that point where regression line touches Y axis. In another words distance between the touching point of regression line on Y axis from the point of origin is a. If correlation is +ve regression line touches Y axis above point of origin and in case of –ve regression line touches Y axis below point of origin.

$$x = a + b.y$$

$\therefore$ $$\Sigma x = n.a + b.\Sigma y$$

or $$na = \Sigma x - b\Sigma y$$

both side divided by n.

$$\frac{na}{n} = \frac{\Sigma x}{n} - \frac{b\Sigma y}{n}$$

or $$a = \bar{x}\ b\bar{y}$$

Likewise $$y = a + bx$$

or $$a = y + b\bar{x}$$

$\bar{x}$ is the mean of x series and $\bar{y}$ is the mean of y series.

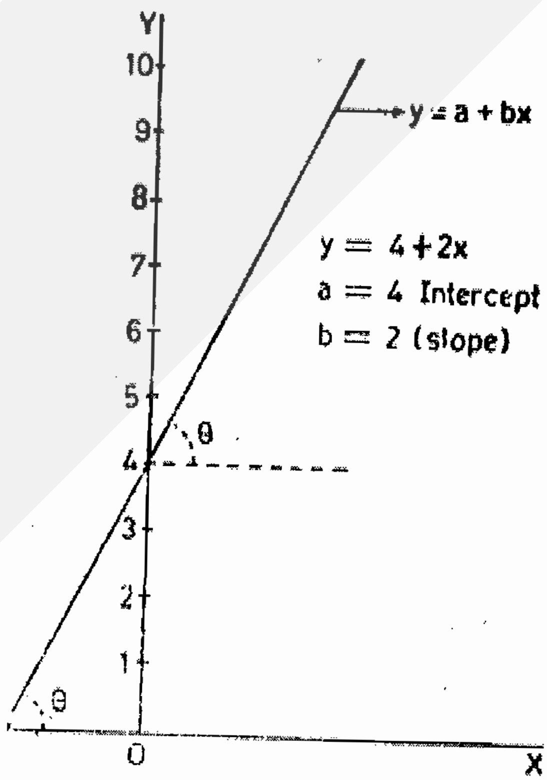


Fig. 10.1. Position of a and b from the equation $y = a + bx$.

Constant b exhibits the slope of the line. It is value of angle made by regression line and its horizontal line (X-axis). In other words b is gradient or slope. It means for the measurement of any distance on X axis :

$$\frac{\text{Change in values of Y axis}}{\text{Distance on axis}}.$$

In the given graph figure position of a and b has been made clear from the equation $y = a + bx$.

This clears that determination of any special straight lines depends on value of a and b and best least square line can be obtained only when real value of a and b is determined. Values of a and b can be obtained by following two normal equations.

In $Y = a + bx$, value of a and b can be obtained by following equation :

$$\Sigma Y = na + b.\Sigma x$$
$$\Sigma XY = a\Sigma x + b\Sigma x^2.$$

Likewise

Normal equation for $x = a' + b'y$, is as follows

$$\Sigma x = na' + b'\Sigma y$$
$$\Sigma xy = a'\Sigma y + b'\Sigma y^2.$$

Regression coefficient. Prediction of the value of one character from the knowledge of the other character is also done by finding another constant called regression coefficient. The significance of the regression coefficient to know whether there is a linear relationship between x and y at the population level tested by t analysis.

Regression coefficient of Y on X is denoted as ***bxy***.

Regression coefficient of Y for one unit of X, and of X for one unit of Y are found by either of the following 3 formulae :

(a) **If correlation coefficient (r) is known,** regression coefficient is derived :

Regression coefficient of X on Y.

$$Bxy = r \times \frac{\text{S.D. of X series}}{\text{S.D. of Y series}} \quad \text{or} \quad r \times \frac{\sigma x}{\sigma y}.$$

Regression coefficient of Y on X.

$$Byx = r \times \frac{\text{S.D. of Y series}}{\text{S.D. X series}} \quad \text{or} \quad r \times \frac{\sigma y}{\sigma x}$$

By multiplying both

$$Bxy \times byx = r \times \frac{\sigma x}{\sigma y} \times r \times \frac{\sigma y}{\sigma x}$$

or $$Bxy \times Byx = r^2$$

or $$\sqrt{Bxy \times Byx} = r.$$

It means square root of the product of the Regression coefficient of x on y and Regression coefficient of y on $x = r$.

(*b*) If means are already calculated, the regression coefficients are :

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\Sigma(Y-\overline{Y})^2}$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\Sigma(X-\overline{X})^2}$$

[This is indirect and stenuous method]

(*c*) If means are not to be calculated a simple and direct method is adopted as below :

$$b_{xy} = \frac{\Sigma XY - \frac{\Sigma X.\Sigma Y}{n}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}}$$

$$b_{yx} = \frac{\Sigma XY - \frac{\Sigma X.\Sigma Y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$$

[**Note.** X and Y are the original measurements.]

Example. The body length and head length of 7 fishes of a species Macroghathus aculeatus is as follows

Body length (X) : 13.4, 15.1, 15.3, 16.8, 17.5, 19.2, and 21.2

Head length (Y) : 2.1, 2.3, 2.3, 2.6, 2.7, 3.0, 3.3

Find out regression equation.

Calculation. Following table is prepared with the help of the given data which satisfied the requirements of the formula.

Table 10.1.

X	Y	$X.Y$	X^2
13.4	2.1	28.14	179.56
15.1	2.3	34.73	228.01
15.3	2.3	35.19	234.09
16.8	2.6	43.68	282.24
17.5	2.7	47.25	306.25
19.2	3.0	57.6	368.64
21.2	3.3	69.96	449.44
$\Sigma X = 118.5$	$\Sigma Y = 18.3$	$\Sigma X.Y = 316.55$	$\Sigma X^2 = 2048.23$

$$y = a + bx \; ; xy = xa + bx^2$$

Putting the values in formula

$$18.3 = 7a + 118.5b \quad \text{...(i)}$$

$$316.55 = 118.5a + 2048.23b \quad \text{...(ii)}$$

Multiply equation (*i*) with 118.5

$$18.3 \times 118.5 = 118.5a + 2048.23b.$$

$$2168.55 = 118.5a + 14042.25b \quad \text{...(iii)}$$

Subtracting equation (*ii*) by (*iii*)

$$(2168.55 - 316.55) = 0 + (14042.25 - 2048.23)b$$

$$\therefore \quad b = \frac{2168.55 - 316.55}{14042.25 - 2048.23}$$

$$b = 0.16$$

Now put the value of b in equation (*i*)

$$18.3 = 7a + .16 \times .118.5$$

or $$18.3 = 70 + 18.96$$

or $$7a = 18.3 - 18.96$$

$$= -.6718$$

$$= \frac{-0.66}{7}$$

$$= .0959.$$

Required Regression equation

$$y = -.0959 + .1601x. \quad \textbf{Ans.}$$

Example 2. The body length (X axis) and length between snout to dorsal fin of 7 group of fishes of a species (*Macrognathus aculeatus*) is given below. Find the Regression equation.

Calculation. Make following table which can satisfy the requirement of formula.

Table 10.2.

X	Y	$X.Y$	X^2	*Estimated value of y,* $y' = .15944 + .65134x$	*Error* $y = y'$
13.4	8.9	119.26	179.56	8.8874	.0216
15.1	10.0	151.00	228.01	9.9947	.0053
15.3	10.1	154.53	234.09	10.1249	−.0249
16.8	11.1	186.48	282.24	11.1019	−.0019
17.5	11.6	203.0	306.25	11.5579	.0421
19.2	12.6	241.92	368.64	12.6652	−.0652
21.2	14.0	296.80	449.44	13.9679	.0321
$\Sigma X =$ 118.5	$\Sigma Y =$ 78.3	$\Sigma XY =$ 1352.99	$\Sigma X^2 =$ 2048.23	$\Sigma = 78.2999$	

Regression equation $\quad y = a + bx.$

With the help of normal equations value of constant a and b was deduced by combined equation method :

$$\Sigma y = n.a + b.\Sigma x$$

$$\Sigma xy = a.\Sigma x + b.\Sigma x^2$$

$$78.3 = 7a + 118.5b \qquad ...(i)$$

$$1352.99 = 118.5a + 2048.23b \qquad ...(ii)$$

Divide equation (i) with 7 and equation (ii) with 118.5

$$11.185714 = a + 16.928571b \qquad ...(iii)$$

$$11.417637 = a + 17.284641b \qquad ...(iv)$$

Equation (iv) is subtracted by equation (iii) to obtain value of a and b.

$$11.185714 = a + 16.928571b$$

$$-11.417637 = -a \pm 17.284641b$$

$$-.231923 = -.35607b \qquad ...(v)$$

Multiply (–) both sides of equation and get +ve value

$$.231923 = .35607b$$

or $$.35607b = .231923$$

or $$b = \frac{.231923}{.35607} = .651341.$$

Now put value of b in equation (i) or (ii).

Putting value of b in equation (i)

$$78.3 = 7a + 118.5 \times .651341$$

or $$7a = 78.3 - 77.183908$$

or $$7a = 1.116092$$

or $$a = \frac{1.116092}{7} = .15944.$$

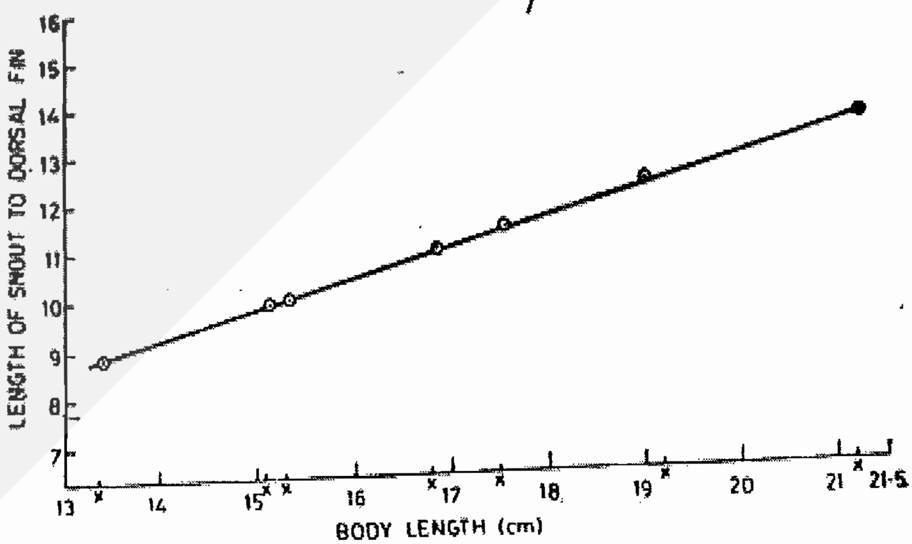


Fig. 10.2. Regression line showing correlation between **body length** and length between snout to dorsal fin of a specie of fish *M. aculeatus*.

Required regression equation is as follows :

$$Y' = .15944 + .651341x.$$

A graphical representation of regression line can be plotted on the basis of data of above table 10.2.

Calculation of regression equation from values of deviation of mean of X and Y variable.

Calculation for regression equation can be simplified by deviations from earn of variables x and y.

In this condition :

$$Y = a + bx$$

equation is represented by

$$Y - \overline{Y} = b(X - \overline{X})$$

or

$$y = bx$$

Here $y = Y - \overline{Y}$ and $x = X - \overline{X}$.

Value of b can be calculated by following formula :

$$b = \frac{\Sigma xy}{\Sigma x^2}.$$

Both normal equations can be represented in terms of X and Y as follows :

$$\Sigma y = n.a + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

$$\because \quad \Sigma y = 0 \text{ and } \Sigma x = 0$$

$$\therefore \quad 0 = n.a + b.\text{zero}.$$

Following table is prepared with the help of data of previous example.

Table 10.3.

X	Y	$X - \overline{X} = x$	$Y - \overline{Y} = y$	$x.y$	x^2	y^2
13.4	8.9	−3.528	−2.285	8.0614	12.446	5.221
15.1	10.0	−1.185	−1.185	2.1662	3.341	1.404
15.3	10.1	−1.085	−1.085	1.7664	2.650	1.177
16.8	11.1	−0.085	−0.085	0.0109	0.016	0.007
17.5	11.6	0.415	0.415	0.2374	0.327	0.172
19.2	12.6	1.415	1.415	3.2149	5.162	2.002
21.2	14.0	2.815	2.815	12.0257	18.249	7.924
$\Sigma X =$ 118.5	$\Sigma Y =$ 78.3			$\Sigma xy =$ 27.4829		$\Sigma y^2 =$ 17.908

Mean $\quad \overline{X} = \frac{118.5}{7} = 16.928$;

Mean $\overline{Y} = \dfrac{78.3}{7} = 11.185.$

Regression equation obtained by following formula :

Y on X

$$Y - \overline{Y} = b(X - \overline{X})$$
$$y = b.x$$
$$\Sigma xy = b.\Sigma x^2$$
$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{27.4829}{42.194} = .6513.$$

Regression equation Y on X

$$Y - 11.185 = .6513\,(X - 16.928)$$

or $$Y - 11.185 = .6513X - .6513 \times 16.928$$

or $$Y - 11.185 = .6513X - 11.0252$$

or $$Y = .6513X - 11.0252 + 11.185$$

or $$Y = .6513X - .159794$$

or $$Y = -0.159794 + .6513X$$

Regression equation X on Y

$$X - \overline{X} = b'\,(Y - \overline{Y})$$
$$x = b'.y \qquad [\because\ X - \overline{X} = x \text{ deviation},\ Y - \overline{Y} = y \text{ deviation}]$$

$\therefore$ $$b' = \frac{\Sigma xy}{\Sigma y^2} \quad \text{or} \quad b' = \frac{27.4829}{17.908} = 1.5347.$$

Regression equation X on Y.

$$X - 16.928 = 1.5347\,(Y - 11.185)$$
$$X - 16.928 = 1.5347Y - 1.5347 \times 11.185$$

or $$X - 16.928 = 1.5347Y - 17.166$$

or $$X = 1.5347Y - 17.166 + 16.928$$

or $$X = 1.5347Y - 0.2376$$

or $$X = -0.2376 + 1.5347Y.$$ **Ans.**

Standard deviation for the regression line :

The variability of the measurements from a regression line may be summarised by standard deviation.

For $$X \text{ S.D. reg.} = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n}}.$$

For $$Y \text{ S.D. reg.} = \sqrt{\frac{\Sigma(Y - \overline{Y})^2}{n}}.$$

Here. Putting the above data.

For $$X \text{ S.D. reg.} = \sqrt{\frac{42.194}{7}} = \sqrt{6.0277142} = 2.455$$

For Y S.D. reg. $= \sqrt{\frac{17.908}{7}} = \sqrt{2.558} = 1.599$

Regression coefficient

$$\text{X on Y or B}xy = r\frac{\sigma x}{\sigma y}$$

$$\text{Y on X or B}yx = r\frac{\sigma y}{\sigma x}.$$

Calculation of regression coefficient X on Y

$$X = -.2376 + 1.3474\,Y$$
$$\sigma x = 2.455$$
$$\sigma y = 1.599$$

$$Bxy = r\frac{\sigma x}{\sigma y}$$

$$1.5347 = r\frac{2.455}{1.599}$$

or $$1.5347 = 1.5353r$$

$\therefore$ $$r = \frac{1.5347}{1.5353} = .9996$$

To find out value of 'r' with the help of both regression equation.

$$Bxy = 1.5347$$

and $$Byx = 0.6513$$

$\therefore$ $$r^2 = Bxy \times Byx$$

or $$r = \sqrt{Bxy \times Byx}$$
$$= \sqrt{1.5347 \times .6513}$$
$$= \sqrt{.9995}$$
$$= 0.999.$$ **Ans.**

It means the **length of body** and length between snout to dorsal fin of *M. aculeatus* has complete positive correlation.

Example. Body length and body weight of 25 *Anabas scandens*, were obtained for a random sample of 25 fish from a pond. The data was obtained as follows. Calculate the coefficient and regression equation.

$$\frac{\text{Body length (cm)}}{\text{Body weight (g)}} \quad \frac{40}{7}, \frac{45}{7.5}, \frac{55}{8.3}, \frac{41}{6.2}, \frac{35}{5.5}, \frac{37}{6}, \frac{50}{8}, \frac{51}{8.2}, \frac{55}{9.2}, \frac{60}{9.5}, \frac{30}{5.2},$$

$$\frac{32}{5.5}, \frac{65}{9.8}, \frac{68}{10.3}, \frac{49}{8}, \frac{52}{8.5}, \frac{40}{5.8}, \frac{43}{6.1}, \frac{37}{5.1}, \frac{39}{5.2}, \frac{52}{8}, \frac{65}{10.1}, \frac{41}{7.9}, \frac{45}{7.8}, \frac{38}{6.2},$$

Arrange the data in a tabular form as shown below :

Table 10.5.

X	Y	X^2	Y^2	XY
40	7.0	1600	49.00	280.0
45	7.5	2025	56.25	337.5
55	8.3	3025	68.89	456.5
41	6.2	1681	38.44	254.2
35	5.5	1225	30.25	192.5
37	6.0	1369	36.00	222.0
50	8.0	2500	64.00	400.0
51	8.5	2601	72.25	433.5
55	9.2	3025	84.64	506.0
60	9.5	3600	90.25	570.0
30	5.2	900	27.04	156.0
32	5.5	1024	30.25	176.0
65	9.8	4225	96.04	637.0
68	10.3	4624	106.09	700.0
49	8.0	2401	64.00	392.0
52	8.5	2704	72.25	442.0
40	5.8	1600	33.64	232.0
43	6.1	1849	37.21	262.3
37	5.1	1369	26.01	188.7
39	5.2	1521	27.04	202.8
52	8.0	2704	64.00	416.0
65	10.1	4225	102.01	656.5
41	7.9	1681	62.41	323.9
45	7.8	2025	60.84	351.0
38	6.2	1444	38.44	235.6
$\Sigma X = 1165$	$\Sigma Y = 185.20$	$\Sigma X^2 = 56947$	$\Sigma Y^2 = 1437.24$	$\Sigma X.Y = 9024.40$

$$\overline{X} = \frac{1165}{25} = 46.25$$

$$\overline{Y} = \frac{185.2}{25} = 7.41$$

$$\Sigma dx^2 = 56947 - \frac{(1165)^2}{25}$$

$$= 56947 - 54289 = 2658.0$$

$$\Sigma dy^2 = 1437.24 - \frac{(185.2)^2}{25}$$

$$= 1437.24 - 1371.96 = 65.28$$

$$\Sigma dxdy = 9024.4 - \frac{1165 \times 185.2}{25}$$

$$= 9024.4 - 8630.32 = 394.08$$

$$r = \frac{\Sigma dxdy}{\sqrt{\Sigma dx^2 \Sigma dy^2}} = \frac{394.08}{\sqrt{2658 \times 65.28}}$$

$$= \frac{394.08}{\sqrt{173514.24}} = \frac{394.08}{416.5} = 0.9461 \quad \text{or} \quad 0.95$$

$$= 0.95, df = 24, p = 0.001$$

$$\text{SE of } r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{(1 - 0.95)^2}{23}}$$

$$= \sqrt{\frac{(1 - 0.9025)}{23}}$$

$$= \sqrt{\frac{0.0975}{23}}$$

$$= \sqrt{0.00424} = 0.065$$

$$= \frac{r}{\text{SE or } r} = \frac{0.95}{0.065}$$

$$= 14.62 \text{ for } df\ 23$$

Significant level, $p = 0.001$.

Tabulated $t = 3.76$ at $p = 0.001$ and with 23 df.

$$\text{Regression } b = \frac{\Sigma dx.dy}{\Sigma dx^2}$$

$$= \frac{394.08}{2658} = 0.15$$

$$a = \bar{y} - b\bar{x} = 7.41 - b(46.6) = 7.41 - 0.15(46.6)$$

$$= 7.41 - 6.91 = 0.5$$

$$Y_x = 0.5 + 0.15x.$$

Conclusion. Calculated $r = 0.95$

Degree of freedom $25 - 1 = 24$

Tabulated $r = 0.597$ ($p = 0.001$).

The mole correlation coefficient 'r' is above the tabulated value of r at $p = 0.001$. Therefore variables, length and weight of fish are highly correlated.

EXERCISE

1. Define and explain correlation and correlation coefficient with examples.
2. The body weight (g) and intake of oxygen (VO_2 cc./kg/h) of a species of 40 fishes (procured randomly from a pond) was measured and following results were obtained. Find correlation between these two variables.

$$\frac{\text{Body weight } (g)}{\text{Oxygen consumption}} \quad \frac{28.4}{66}, \frac{28.5}{67}, \frac{28.6}{65}, \frac{28.7}{73}, \frac{28.8}{74}, \frac{28.9}{75}, \frac{29.1}{83}, \frac{29.2}{84}, \frac{29.3}{85},$$

$$\frac{32.1}{84}, \frac{32.2}{85}, \frac{32.3}{86}, \frac{30.8}{99}, \frac{30.9}{100}, \frac{31}{101}, \frac{30.5}{104}, \frac{30.6}{105}, \frac{30.7}{106}, \frac{27.7}{94}, \frac{27.8}{95}, \frac{27.9}{96}, \frac{27.2}{89},$$

$$\frac{27.3}{90}, \frac{27.4}{91}, \frac{29.0}{89}, \frac{29.1}{90}, \frac{29.2}{91}, \frac{29.3}{77}, \frac{29.4}{78}, \frac{29.5}{79}, \frac{27.8}{73}, \frac{27.9}{74}, \frac{28}{75}, \frac{31.3}{63}, \frac{31.4}{64},$$

$$\frac{31.5}{65}, \frac{26.4}{60}, \frac{26.5}{61}, \frac{26.6}{62}, \frac{26.7}{63}. \qquad N = 40$$

3. Deviation taken from mean of X and Y (two variable) are given below. Find 'r' by Pearson's product moment method and explain their significance.
   X – 4, –3, –2, –1, 0, 1, 2, 3, 4
   Y – 3, –3, –4, 0, 4, 4, 1, 2, –2, –1.
4. Data of few pair of different age groups (husband and wife's age) is given below. Find correlation and explain the significance.

Age of husband	*Age of wife*					
	10–20	*20–30*	*30–40*	*40–50*	*50–60*	*Sum*
10–20	6	3	—	—	—	9
20–30	3	16	10	—	—	29
30–40	—	10	15	7	—	32
40–50	—	—	7	10	4	21
50–60	—	—	—	4	5	9
Sum	9	29	32	21	9	100

5. Two Judges gave following ranks to 12 entries in a baby show. Find the rank correlation coefficient.

Entries	*Judge I*	*Judge II*
1	7	6
2	8	4
3	2	1
4	1	3
5	9	11
6	3	2
7	12	12
8	11	10
9	4	5
10	10	9
11	6	7
12	5	8

6. An experimenter tried to establish relationship between Earthworm density and soil pH. 15 quadrats of size 25 × 25 cm were laid randomly. The results are as follows :

Quadrat No.	*Soil pH*	*No. of Earthworm*
1	6.8	15
2	7.2	20
3	7.0	18
4	6.9	22
5	7.5	18
6	6.9	19
7	7.1	25
8	7.4	20
9	7.2	21
10	7.5	18
11	6.5	16
12	6.8	20
13	7.3	19
14	6.9	21
15	7.4	18

7. What is regression ? Differentiate between correlation and regression. Explain the methods of least square to estimate the regression coefficient in a linear regression of Y on X.

8. Calculate the 'r' between two measurements of water quality of a lake which are given below. Show the level of significance of 'r'. Calculate the Y-intercept a and regression coefficient b.

Salinity %	5	7	9	3	16	14
Dissolved O_2 (mg^{1-1})	7	5	5	9	3	2

9. What is the purpose of regression analysis ? What do you mean by linear regression. Explain regression equation.

10. The body length and girth of 7 groups of a species of fish in cm is as follows. Find the regression equation.

Body length—X	*Girth of body—Y*
13.9	4.2
15.7	4.7
15.8	4.7
17.5	5.2
18.1	5.4
19.9	6.0
22.0	6.5

Standard Formulae

MEAN

Ungrouped data :

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{N}$$

or

$$\bar{X} = \frac{\Sigma X}{N} \quad \text{or} \quad \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Grouped data :

$$\bar{X} = \frac{f_1 . X_1 + f_2 . X_2 + f_3 . X_3 \ldots + f_n . X_n}{f_1 + f_2 + f_3 + \ldots + f_n} \quad \text{or} \quad \frac{\Sigma f.m}{\Sigma f}$$

Weighted arithmetic mean :

$$\bar{X}_w = \frac{W_1 . X_1 + W_2 . X_2 + W_3 . X_3 + \ldots + W_n . X_n}{W_1 + W_2 + W_3 + \ldots + W_n}$$

Geometric mean :

$$(GM) = n\sqrt{x_1 . x_2 . x_3 \ldots x_n} = (x_1 . x_2 . x_3 \ldots . x_n)^{1/n}$$

$$GM = \text{Antilog}\,(\Sigma f . \log x / \Sigma f)$$

This can be written as

$$GM = \text{antilog}\left(\frac{1}{n}\sum_{i=1}^{n} \log x\, i\right)$$

Weighted geometric mean :

$$GM = N\sqrt{X_1^{W_1} \times X_2^{W_2} \times \ldots \times X_n^{W_n}} \quad \text{where, } N = W_1 + W_2 + \ldots + W_n$$

$$\therefore \quad \log GM = \frac{w_1 . \log x_1 + w_2 . \log x_2 + \ldots + w_n . \log x_n}{w_1 + w_2 + \ldots + w_n}$$

$$= \frac{\Sigma w_i \log x_i}{\Sigma w_i}$$

Harmonic Mean :

$$HM = \frac{N}{1/x_1 + 1/x_2 + 1/x_3 + \ldots + 1/x_n}$$

or

$$HM = \frac{N}{\Sigma\,(1/x)}$$

$$\text{HM} = \frac{f_1 + f_2 + f_3 + \ldots + f_n}{f_1/X_1 + f_2/X_2 + f_3/X_3 + \ldots + f_n/X_n}$$

or,
$$\text{HM} = \frac{\Sigma f}{\Sigma (f/X)}.$$

MEDIAN

(i) Median value is the value of the (N + 1/2)th item

(when N is odd) $\text{Median} = l_1 + \dfrac{(m - F)}{fm} \times (l_2 - l_1)$ (Discrete Series)

(ii) $\text{Median} = \dfrac{(N/2)\text{th} + (N/2 + 1)th}{2}$ item. $\text{Median} = l_1 + \dfrac{(\Sigma f/2 - F)}{fm} \times i$

(when N is even) (Continuous series)

MODE

Mode of a frequency distribution is defined as *"that value of the variable for which the frequency is maximum."*

Mode (*for a discrete series*) : If the distribution is regular and only one maximum frequency is there (data is a unimodal) then the mode value can be obtained by mere inspection.

Sometimes a series have more than one mode (bimodal or multimodal). Then mode is obtained by grouping method.

Mode (*for continuous series*) : In the case of bimodal or trimodal condition we prepare grouping and analysis table and find out the modal class. Then apply the formula.

$$\text{Mode} = L_1 + \frac{f_1 + f_0}{f_1 - f_0 - f_2} \times (L_2 - L_1) \text{ or } L_1 + \left[\frac{\Delta_1}{\Delta 1 + \Delta_2}\right] \times i$$

Empirical formula for mode : Mode is also computed by the empirical relation.

Mode = 3 Median – 2 Mean

QUARTILES

To divide the total frequency into four equal parts, three partition values are essential and they are denoted by Q_1, Q_2 and Q_3. Here Q_1 is called the **lower or first quartile** and Q_3 the **upper or third quartile.**

Quartiles for an individual series : Quartiles are located by inspection method. The series is arranged in ascending order of magnitude and following formula is used to compute quartiles.

$$Q_1 = \text{size of } \left[\frac{n+1}{4}\right]\text{th item}$$

$$Q_2 = \text{size of } 2\left[\frac{n+1}{4}\right]\text{th item}$$

$$Q_3 = \text{size of } 3\left[\frac{n+1}{4}\right]\text{th item.}$$

PERCENTILES

Values dividing a series (arranged in ascending or descending order of magnitude) in hundred equal parts. There are, therefore, 99 percentiles denoted by $P_1, P_2, P_3, P_4, \ldots, P_{99}$. **50th percentile is median** *i.e.* $P_{50} = Q_2$.

The percentiles are computed (*i*) For ungrouped data and (*ii*) grouped data (discrete series) by using formula :

$$P_i = l_1 + hX\,(iN/100 - fc)/fm$$

RANGE

Range = H – L,

Here R = Range, H = Highest value of variable, L = Lowest value of variable.

$$\text{Coefficient of range} = \frac{L - S}{L + S} = \frac{\text{Difference of extreme items}}{\text{Sum of extreme items}}.$$

QUARTILE DEVIATION

Grouped data :

$$Q_1 = L + \frac{(f/4 - F)}{fq} \times i \; ; \; Q_3 = L + \frac{(3f/4 - F)}{fq} \times i$$

$$Q = \frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{2} = \frac{Q_3 - Q_1}{2}$$

Here, L = Lower limit of that class interval, where $Q_1\left(\frac{N}{4}\right)$ or $Q_3\left(\frac{3N}{4}\right)$ falls.

F = Cumulative frequency just above that class interval where

$$Q_1\left(\frac{N}{4}\right) \quad \text{or} \quad Q_3\left(\frac{3N}{4}\right) \text{ falls.}$$

fq = frequency of that class interval where Q_1 and Q_3 falls ; i = length of class interval.

Co-efficient of quartile deviation :

$$\text{Co-efficient of Q} = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

MEAN DEVIATION

Mean Deviation or MD or $\delta = \frac{\Sigma|x|}{N}, \quad \delta = \frac{\Sigma|fx|}{\Sigma f}$

Co-efficient of mean deviation : Co-efficient of mean deviation denote by C of $\delta\,\overline{X}$ and it is obtained by formula :

$$C \text{ of } \delta\,\overline{X} = \frac{\delta\,\overline{X}}{\overline{X}}.$$

STANDARD DEVIATION

$$S = \frac{\sqrt{\Sigma x^2}}{N} \quad \text{or} \quad S = \frac{\sqrt{\Sigma x^2}}{N-1}; \quad S = \frac{\sqrt{\Sigma fx^2}}{\Sigma f} \quad \text{or} \quad S = \frac{\sqrt{\Sigma f\,.\,x^2}}{\Sigma f - 1}$$

Computation of standard deviation by direct method :

$$S = \sqrt{\frac{\Sigma fX^2 - \overline{X}^2}{\Sigma f}}.$$

Co-efficient of standard deviation :

$$\text{Coefficient of } S = \frac{S}{\overline{X}}$$

Combined standard deviation : If two frequency distributions have means $\overline{X}_1$ and $\overline{X}_2$ and standard deviations σ_1 and σ_2 respectively. Then, the combined S.D., denoted by σ_{12}, of the two distributions is obtained by using :

$$\sigma_{12} = \frac{\sqrt{N_1(\sigma_1^2 + D_1^2) + N_2(\sigma_2^2 + D_2^2)}}{N_1 + N_2}$$

The above formula can be extended to any number of distributions. For example, in the case of three distributions, it will be :

$$\sigma_{123} = \frac{\sqrt{N_1(\sigma_1^2 + D_1^2) + N_2(\sigma_2^2 + D_2^2) + N_3(\sigma_3^2 + D_3^2)}}{N_1 + N_2 + N_3}$$

$$SE\overline{X} = \frac{SD}{\sqrt{N}} \text{ or } \frac{S}{\sqrt{N}}$$

$$SD \text{ or } S \text{ or } \sigma = \sqrt{\frac{\Sigma f x^2}{\Sigma f}}$$

(In case number of observations are more than 30)

$$SD \text{ or } S \text{ or } \sigma = \frac{\sqrt{\Sigma f x^2}}{\Sigma f - 1}$$

(In case number of observations are less than 30)

Standard Error of the Standard Deviation :

$$SE_\sigma = \frac{SD}{\sqrt{2N}} \text{ or } \frac{\sigma}{\sqrt{2N}} \qquad SE_\sigma = \frac{SD}{\sqrt{2.\Sigma f}}$$

Z TEST

To test the significance of difference between two sample means or between experiment sample mean and a control sample mean.

$$Z = \frac{\text{observed difference between two sample means}}{\text{SE of difference between two sample means}}$$

or

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sigma(\overline{X}_1 - \overline{X}_2)}$$

Standard error of difference is denoted as $S(\overline{X}_1 - X_2)$ or $\sigma(X_1 - \overline{X}_2)$. It is computed by a formula

$$\sigma(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

STUDENT'S 't' TEST

$$'t' = \frac{\overline{X} \times \sqrt{N}}{SD} \quad \text{(Unpaired data)}$$

$$t = \frac{\overline{X}_1 - X_2}{SE_D} \quad \text{(Paired data)}$$

and

$$SE_D = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

WEB: www.evolution.co.in | MAIL: info@evolution.co.in | PHONE: +91 11 4709 2329 ; + 91 9810 975 937 | ©Evolution Educare Pvt. Ltd.

THE CHI-SQUARE TEST

$$\chi^2 = \Sigma\left\{\frac{(0-E)^2}{E}\right\} \quad \text{or} \quad \Sigma\left\{\frac{(f_0-f_e)^2}{f_e}\right\}$$

PROBABILITY

Statistically, probability can be explained in the following way. If an event can happen in 'a' ways, and fail to happen in 'b' ways, then the probability of its happening 'p' can be written as :

$$p = \frac{a}{a+b} \quad \text{or} \quad p = \frac{\text{Number of events occurring}}{\text{Total number of trials}}.$$

Similarly, the probability of the failure of the event to happen is denoted by 'q'. Therefore,

$$q = \frac{b}{a+b}$$

$$\therefore \quad p+q = \frac{a}{a+b} + \frac{b}{a+b} = \frac{a+b}{a+b} = 1$$

$$\because \quad p+q = 1 \quad \therefore \quad p = 1-q \text{ and } q = 1-p.$$

Addition rule of probability : This rule is applied when events are mutually exclusive *i.e.*, both events cannot occur simultaneously.

Mathematically, if $P(E_1)$ and $P(E_2)$ are the respective probability of two mutually exclusive events E_1 and E_2, then the probability of happening of any one can be expressed as follows :

$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2).$$

The rule can be extended to any number of mutually exclusive events as follows :

$$P(E_1 \text{ or } E_2 \text{ or } E_3 \ldots \text{ or } E_n) = P(E_1) + P(E_2) + P(E_3) + \ldots + P(E_n)$$

CORRELATION

Karl Pearson's Coefficient of correlation is given by the expression :

$$r_{x,y} = \frac{\Sigma x \,.\, y}{\sqrt{\Sigma x^2 \,.\, \Sigma y^2}}$$

Calculation of 'r' by using raw scores :

$$r = \frac{N\Sigma X \,.\, Y - \frac{\Sigma X \,.\, \Sigma Y}{N}}{\sqrt{N\,[\Sigma X^2 - (\Sigma X)^2] \times N\,[\Sigma Y^2 - (\Sigma Y)^2]}}$$

Verification of significance of correlation coefficient :

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-(r)^2}}.$$

Rank correlation :

$$\rho = 1 - \frac{6\Sigma D^2}{n(n^2-1)}$$

where ρ (Rho) is the rank difference of X and Y variables, D is the difference between the pair of the same individual in the two characteristics and n is number of pairs. ΣD^2 is summation of square of difference of two variables rank I and II (R_1 and R_2).

Correlation between three variables :

$$r_{123} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

where r_{12}, r_{13} and r_{23} are the correlation coefficient between the pairs of variates ; X_1 and X_2 ; X_1 and X_3 ; X_2 and X_3 respectively.

REGRESSION-EQUATIONS

Regression equation of X on Y : $xy = a + by$.

Regression equation of Y on X : $yx = a + bx$.

Here X and Y are the variables whereas 'a' and 'b' are unknown constants. The constant 'a' is the distance between the point of origin and the points where the regression line touches the Y axis. The constant 'b' shows the slope of the line and is also called regression coefficient. It indicates that for every one unit change in X, there will be two units changes in Y.

The constant 'a' and 'b' can be obtained by the following formula :
$a = \overline{Y} - b\,\overline{X}$ (where $\overline{X}$ and $\overline{Y}$ are their respective means).

$$b = \frac{\Sigma x\,.\,y}{\Sigma x^2} = \frac{\Sigma X\,.\,Y - \frac{\Sigma X\,.\,\Sigma Y}{N}}{\Sigma x^2 - \frac{(\Sigma X)^2}{N}}$$

where $x = (X - \overline{X})$ and $y = (Y - \overline{Y})$.

Regression coefficient of Y for one unit of X and regression coefficient of X for one unit of Y are found by either of the following 3 formulae :

(i) If correlation coefficient (r) is known, regression coefficient is derived as :

Regression coefficient of X on Y.

$$b_{xy} = r \times \frac{\sigma \text{ of X series}}{\sigma \text{ of Y series}} \text{ or } r \times \frac{\sigma X}{\sigma Y}$$

WEB: www.evolution.co.in | MAIL: info@evolution.co.in | PHONE: +91 11 4709 2329 ; + 91 9810 975 937 | ©Evolution Educare Pvt. Ltd.

Regression coefficient of Y on X.

$$b_{yx} = r \times \frac{\sigma \text{ of Y series}}{\sigma \text{ of X series}} \text{ or } r \times \frac{\sigma Y}{\sigma X}$$

Multiplying both equations

$$b_{xy} \cdot b_{yx} = r \times \frac{\sigma X}{\sigma Y} \times r \times \frac{\sigma Y}{\sigma X}$$

or $\quad b_{xy} \cdot b_{yx} = r^2$

or $\quad \sqrt{b_{xy} \cdot b_{yx}} = r$

Ir means square root of the product of the regression coefficient of X on Y and regression coefficient of Y on X = r.

(*ii*) If means are already calculated, the regression coefficient are

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} \text{ or } \frac{\Sigma (X - \overline{X})(Y - \overline{Y})}{\Sigma (Y - \overline{Y})^2}$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} \text{ or } \frac{\Sigma (X - \overline{X})(Y - \overline{Y})}{\Sigma (X - \overline{X})^2}$$

[The above method is strenuous and indirect].

If means are not to be calculated a simple and direct method is adopted as below :

(*iii*)
$$b_{xy} = \frac{\Sigma XY - \dfrac{\Sigma X \cdot \Sigma Y}{N}}{\Sigma Y^2 - \dfrac{(\Sigma Y)^2}{N}}$$

$$b_{yx} = \frac{\Sigma XY - \dfrac{\Sigma X \cdot \Sigma Y}{N}}{\Sigma X^2 - \dfrac{(\Sigma X)^2}{N}}$$

VITAL STATISTICS

Crude rate :

$$= \frac{\left[\begin{array}{c}\text{Total No. of events that occurred in a given} \\ \text{geographical area during a given period}\end{array}\right]}{\left[\begin{array}{c}\text{Mid year population of the geographical} \\ \text{area for the same period}\end{array}\right]} \times 1000$$

Specific rate :

$$= \frac{\left[\begin{array}{c}\text{No. of events which occurred among a specific}\\ \text{group of the population of a given}\\ \text{geographical area during a given year}\end{array}\right]}{\left[\begin{array}{c}\text{Mid year population of the specific group of the}\\ \text{population in the same geographical area}\\ \text{during the same period}\end{array}\right]} \times 1000$$

Crude birth rate $= \dfrac{\text{Annual births}}{\text{Annual mean population}} \times 1000$

$$\text{S.F.R.} = \frac{\left[\begin{array}{c}\text{Number of live births which occurred to females of a specified}\\ \text{age group of the poulation of a region during a given year}\end{array}\right]}{\left[\begin{array}{c}\text{Mid-year female population of the specified age-group year}\\ \text{in the geographical area during the same year}\end{array}\right]} \times 1000$$

General fertility rate (G.F.R.)

$$= \frac{\left[\begin{array}{c}\text{No. of live births which occurred among the population}\\ \text{of a given region during a given year}\end{array}\right]}{\left[\begin{array}{c}\text{Female population of age 15 to 49 in the}\\ \text{given region during the same year}\end{array}\right]} \times 1000$$

Gross reproductive rate :

$$\text{GRR} = \frac{\text{No. of female births}}{\text{Total no. of births}} \times \text{Total fertility rate}$$

Also $\quad \text{GRR} = \dfrac{\text{No. of female children born 1,000 woman}}{1,000}$

Net reproduction rate. NRR $= Sb \overset{49}{\underset{15}{x}} \dfrac{L_x}{l_0}$

Crude death rate $= \dfrac{\text{Annual deaths}}{\text{Annual mean populaiton}} \times 1000$

Infant mortality rate $= \dfrac{\text{No. of deaths under 1 year of age}}{\text{No. of live birth}} \times 1000$

(*i*) Foetal death ratio $= \dfrac{\text{Foetal death}}{\text{Live birth}} \times 1000$

(*ii*) Perinatal mortality rate

$$= \frac{\text{Late foetal death + deaths one week}}{\text{Total births (Live + still)}} \times 1000$$

(*iii*) Neonatal mortality rate

$$= \frac{\text{No. of deaths upto 28 days of life}}{\text{No. of live births}} \times 1000.$$

LOGARITHMS

											Mean Differences								
	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374	4	8	12	17	21	25	29	33	37
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	11	15	19	23	26	30	34
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	3	7	10	14	17	21	24	28	31
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	6	10	13	16	19	23	26	29
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	9	12	15	18	21	24	27
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	8	11	14	17	20	22	25
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	5	8	11	13	16	18	21	24
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	2	5	7	10	12	15	17	20	22
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	2	5	7	9	12	14	16	19	21
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	2	4	7	9	11	13	16	18	20
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	2	4	6	8	11	13	15	17	19
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	6	8	10	12	14	16	18
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	6	8	10	12	14	15	17
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	7	9	11	13	15	17
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	5	7	9	11	12	14	16
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298	2	3	5	7	8	10	11	13	15
27	4314	4330	4386	4362	4378	4393	4409	4425	4440	4456	2	3	5	6	8	9	11	13	14
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609	2	3	5	6	8	9	11	12	14
29	4624	4639	4654	4669	4683	4648	4713	4728	4742	4757	1	3	4	6	7	9	10	12	13
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900	1	3	4	6	7	9	10	11	13
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038	1	3	4	6	7	8	10	11	12
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172	1	3	4	5	7	8	9	11	12
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302	1	3	4	5	6	8	9	10	12
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428	1	3	4	5	6	8	9	10	11
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551	1	2	4	5	6	7	9	10	11
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670	1	2	4	5	6	7	8	10	11
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786	1	2	3	5	6	7	8	9	10
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899	1	2	3	5	6	7	8	9	10
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010	1	2	3	4	5	7	8	9	10
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117	1	2	3	4	5	6	8	9	10
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222	1	2	3	4	5	6	7	8	9
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325	1	2	3	4	5	6	7	8	9
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425	1	2	3	4	5	6	7	8	9
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522	1	2	3	4	5	6	7	8	9
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618	1	2	3	4	5	6	7	8	9
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	7	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	5	5	6	7	8
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893	1	2	3	4	4	5	6	7	8
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981	1	2	3	4	4	5	6	7	8
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	3	4	5	6	7	8

LOGARITHMS

	0	1	2	3	4	5	6	7	8	9	Mean Differences 1	2	3	4	5	6	7	8	9
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	3	4	5	6	7	8
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	2	3	4	5	6	7	7
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	2	3	4	5	6	6	7
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	2	3	4	5	6	6	7
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	2	3	4	5	5	6	7
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	2	3	4	5	5	6	7
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	2	3	4	5	5	6	7
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	1	2	3	4	4	5	6	7
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	1	2	3	4	4	5	6	7
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	5	5	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	3	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	6
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	3	4

ANTILOGARITHMS

											Δm	1	2	3	4	5	6	7	8	9
x	0	1	2	3	4	5	6	7	8	9	+	ADD								
.00	1000	1002	1005	1007	1009	1012	1014	1016	1019	1021	2	0	0	1	1	1	1	1	2	2
.01	1023	1026	1028	1030	1033	1035	1038	1040	1042	1045	2	0	0	1	1	1	1	1	2	2
.02	1047	1050	1052	1054	1057	1059	1062	1064	1067	1069	2	0	0	1	1	1	1	1	2	2
.03	1072	1074	1076	1079	1081	1084	1086	1089	1091	1094	2	0	0	1	1	1	1	1	2	2
.04	1096	1099	1102	1104	1107	1109	1112	1114	1117	1119	3	0	1	1	1	1	2	2	2	3
.05	1122	1125	1127	1130	1132	1135	1138	1140	1143	1146	3	0	1	1	1	1	2	2	2	3
.06	1148	1151	1153	1156	1159	1161	1164	1167	1169	1172	3	0	1	1	1	1	2	2	2	3
.07	1175	1178	1180	1183	1186	1189	1191	1194	1197	1199	3	0	1	1	1	1	2	2	2	3
.08	1202	1205	1208	1211	1213	1216	1219	1222	1225	1227	3	0	1	1	1	1	2	2	2	3
.09	1230	1233	1236	1239	1242	1245	1247	1250	1253	1256	3	0	1	1	1	1	2	2	2	3
.10	1259	1262	1265	1268	1271	1274	1276	1279	1282	1285	3	0	1	1	1	1	2	2	2	3
.11	1288	1291	1294	1297	1300	1303	1306	1309	1312	1315	3	0	1	1	1	2	2	2	2	3
.12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	3	0	1	1	1	2	2	2	2	3
.13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	3	0	1	1	1	2	2	2	2	3
.14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	3	0	1	1	1	2	2	2	2	3
.15	1413	1416	1419	1422	1426	1429	1432	1435	1439	1442	3	0	1	1	1	2	2	2	2	3
.16	1445	1449	1452	1455	1459	1462	1466	1469	1472	1476	3	0	1	1	1	2	2	2	2	3
.17	1479	1483	1486	1489	1493	1496	1500	1503	1507	1510	4	0	1	1	2	2	2	3	3	4
.18	1514	1517	1521	1524	1528	1531	1535	1538	1542	1545	4	0	1	1	2	2	2	3	3	4
.19	1549	1552	1556	1560	1563	1567	1570	1574	1578	1581	4	0	1	1	2	2	2	3	3	4
.20	1585	1589	1592	1596	1600	1603	1607	1611	1614	1618	4	0	1	1	2	2	2	3	3	4
.21	1622	1626	1629	1633	1637	1641	1644	1648	1652	1656	4	0	1	1	2	2	2	3	3	4
.22	1660	1663	1667	1671	1675	1679	1683	1687	1690	1694	4	0	1	1	2	2	2	3	3	4
.23	1698	1702	1706	1710	1714	1718	1722	1726	1730	1734	4	0	1	1	2	2	2	3	3	4
.24	1738	1742	1746	1750	1754	1758	1762	1766	1770	1774	4	0	1	1	2	2	2	3	3	4
.25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	4	0	1	1	2	2	2	3	3	4
.26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	4	0	1	1	2	2	2	3	3	4
.27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	4	0	1	1	2	2	2	3	3	4
.28	1905	1910	1914	1919	1923	1928	1932	1936	1941	1945	4	0	1	1	2	2	2	3	3	4
.29	1950	1954	1959	1963	1968	1972	1977	1982	1986	1991	4	0	1	1	2	2	2	3	3	4
.30	1995	2000	2004	2009	2014	2018	2023	2028	2032	2037	5	0	1	1	2	2	3	3	4	4
.31	2042	2046	2051	2056	2061	2065	2070	2075	2080	2084	5	0	1	1	2	2	3	3	4	4
.32	2089	2094	2099	2104	2109	2113	2118	2123	2128	2133	5	0	1	1	2	2	3	3	4	4
.33	2138	2143	2148	2153	2158	2163	2168	2173	2178	2183	5	0	1	1	2	3	3	4	4	5
.34	2188	2193	2198	2203	2208	2213	2218	2223	2228	2234	5	1	1	2	2	3	3	4	4	5
.35	2239	2244	2249	2254	2259	2265	2270	2275	2280	2286	5	1	1	2	2	3	3	4	4	5
.36	2291	2296	2301	2307	2312	2317	2323	2328	2333	2339	5	1	1	2	2	3	3	4	4	5
.37	2344	2350	2355	2360	2366	2371	2377	2382	2388	2393	6	1	1	2	2	3	4	4	5	5
.38	2399	2404	2410	2415	2421	2427	2432	2438	2443	2449	6	1	1	2	2	3	4	4	5	5
.39	2455	2460	2466	2472	2477	2483	2489	2495	2500	2506	6	1	1	2	2	3	4	4	5	5
.40	2512	2518	2523	2529	2535	2541	2547	2553	2559	2564	6	1	1	2	2	3	4	4	5	5
.41	2570	2576	2582	2588	2594	2600	2606	2612	2618	2624	6	1	1	2	2	3	4	4	5	5
.42	2630	2636	2642	2649	2655	2661	2667	2673	2679	2685	6	1	1	2	2	3	4	4	5	5
.43	2692	2698	2704	2710	2716	2723	2729	2735	2742	2748	6	1	1	2	2	3	4	4	5	5
.44	2754	2761	2767	2773	2780	2786	2793	2799	2805	2812	6	1	1	2	2	3	4	4	5	5
.45	2818	2825	2831	2838	2844	2851	2858	2864	2871	2877	7	1	1	2	3	3	4	5	6	6
.46	2884	2891	2897	2904	2911	2917	2924	2931	2938	2944	7	1	1	2	3	3	4	5	6	6
.47	2951	2958	2965	2972	2979	2985	2992	2999	3006	3013	7	1	1	2	3	3	4	5	6	6
.48	3020	3027	3034	3041	3048	3055	3062	3069	3076	3083	7	1	1	2	3	4	4	5	6	6
.49	3090	3097	3105	3112	3119	3126	3133	3141	3148	3155	7	1	1	2	3	4	4	5	6	6

ANTILOGARITHMS

x	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
.50	3162	3170	3177	3184	3192	3199	3206	3214	3221	3228	1	1	2	3	4	4	5	6	7
.51	3236	3243	3251	3258	3266	3273	3281	3289	3296	3304	1	2	2	3	4	5	5	6	7
.52	3311	3319	3327	3334	3342	3350	3357	3365	3373	3381	1	2	2	3	4	5	5	6	7
.53	3388	3396	3404	3412	3420	3428	3436	3443	3451	3459	1	2	2	3	4	5	6	6	7
.54	3467	3475	3483	3491	3499	3508	3536	3524	3532	3540	1	2	2	3	4	5	6	6	7
.55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	7
.56	3631	3639	3648	3656	3664	3673	3681	3690	3608	3707	1	2	3	3	4	5	6	7	8
.57	3715	3724	3733	3741	3750	3758	3707	3776	3784	3793	1	2	3	3	4	5	6	7	8
.58	3802	3811	3819	3828	3837	3846	3855	3864	3873	3882	1	2	3	4	4	5	6	7	8
.59	3890	3899	3908	3917	3926	3936	3945	3954	3963	3972	1	2	3	4	5	6	6	7	8
.60	3981	3990	3999	4009	4018	4027	4036	4046	4055	4064	1	2	3	4	5	6	6	7	8
.61	4074	4083	4093	4102	4111	4121	4130	4140	4150	4159	1	2	3	4	5	6	7	8	9
.62	4169	4178	4188	4198	4207	4217	4227	4236	4246	4256	1	2	3	4	5	6	7	8	9
.63	4266	4276	4285	4295	4305	4315	4325	4335	4345	4355	1	2	3	4	5	6	7	8	9
.64	4365	4375	4385	4395	4406	4410	4426	4436	4446	4457	1	2	3	4	5	6	7	8	9
.65	4467	4477	4487	4498	4508	4519	4529	4539	4550	4560	1	2	3	4	5	6	7	8	9
.66	4571	4581	4592	4603	4613	4624	4634	4645	4656	4667	1	2	3	4	5	6	7	9	10
.67	4677	4688	4699	4710	4721	4732	4742	4753	4764	4775	1	2	3	4	5	7	8	9	10
.68	4786	4797	4808	4819	4831	4840	4853	4864	4875	4887	1	2	3	4	6	7	8	9	10
.69	4898	4909	4920	4932	4943	4955	4966	4977	4989	5000	1	2	3	5	6	7	8	9	10
.70	5012	5023	5035	5047	5058	5070	5082	5093	5105	5117	1	2	4	5	6	7	8	9	11
.71	5129	5140	5152	5164	5176	5188	5200	5212	5224	5236	1	2	4	5	6	7	8	10	11
.72	5248	5260	5272	5284	5297	5309	5321	5333	5346	5358	1	2	4	5	6	7	9	10	11
.73	5370	5383	5395	5408	5420	5433	5445	5458	5470	5483	1	3	4	5	6	8	9	10	11
.74	5495	5508	5521	5534	5546	5559	5572	5585	5598	5610	1	3	4	5	6	8	9	10	12
.75	5623	5636	5649	5662	5675	5689	5702	5715	5728	5741	1	3	4	5	7	8	9	10	12
.76	5754	5768	5781	5794	5808	5821	5834	5848	5861	5875	1	3	4	5	7	8	9	11	12
.77	5888	5902	5916	5929	5943	5957	5970	5984	5998	6015	1	3	4	5	7	8	10	11	12
.78	6026	6039	6053	6067	6081	6095	6109	6124	6138	6152	1	3	4	6	7	8	10	11	13
.79	6166	6180	6194	6209	6223	6237	6252	6206	6281	6295	1	3	4	6	7	9	10	11	13
.80	6310	6324	6339	6353	6368	6383	6397	6412	6427	6442	1	3	4	6	7	9	10	12	13
.81	6457	6471	6486	6501	6516	6531	6546	6561	6577	6592	2	3	5	6	8	9	11	12	14
.82	6607	6622	6637	6653	6668	6683	6699	6714	6730	6745	2	3	5	6	8	9	11	12	14
.83	6761	6776	6792	6808	6823	6839	6855	6871	6887	6902	2	3	5	6	8	9	11	13	14
.84	6918	6934	6950	6966	6982	6998	7015	7031	7047	7063	2	3	5	6	8	10	11	13	15
.85	7079	7096	7112	7129	7145	7161	7178	7194	7211	7228	2	3	5	7	8	10	12	13	15
.86	7244	7261	7278	7295	7311	7328	7345	7362	7379	7396	2	3	5	7	8	10	12	13	15
.87	7413	7430	7447	7464	7482	7499	7516	7534	7551	7568	2	3	5	7	9	10	12	14	16
.88	7586	7603	7621	7638	7656	7674	7691	7709	7727	7745	2	4	5	7	9	11	12	14	16
.89	7762	7780	7798	7816	7834	7852	7870	7889	7907	7925	2	4	5	7	9	11	13	14	16
.90	7943	7962	7980	7998	8017	8035	8054	8072	8091	8110	2	4	6	7	9	11	13	15	17
.91	8128	8147	8166	8185	8204	8222	8241	8260	8279	8299	2	4	6	8	9	11	13	15	17
.92	8318	8337	8356	8375	8395	8414	8433	8453	8472	8492	2	4	6	8	10	12	14	15	17
.93	8511	8531	8551	8570	8590	8610	8630	8650	8670	8690	2	4	6	8	10	12	14	16	18
.94	8710	8730	8750	8770	8790	8810	8831	8851	8872	8892	2	4	6	8	10	12	14	16	18
.95	8913	8933	8954	8974	8995	9016	9036	9057	9078	9099	2	4	6	8	10	12	15	17	19
.96	9120	9141	9162	9183	9204	9226	9247	9268	9290	9311	2	4	6	8	11	13	15	17	19
.97	9333	9354	9376	9397	9419	9441	9462	9484	9506	9528	2	4	7	9	11	13	15	17	20
.98	9550	9572	9594	9616	9638	9661	9683	9705	9727	9750	2	4	7	9	11	13	16	18	20
.99	9772	9795	9817	9840	9863	9886	9908	9931	9954	9977	2	5	7	9	11	14	16	18	20

PERCENTILE VALUES (t_p) for STUDENTS t DISTRIBUTION with ν degrees of freedom (shaded area = p)

ν	$t_{.995}$	$t_{.90}$	$t_{.975}$	$t_{.95}$	$t_{.90}$	$t_{.80}$	$t_{.75}$	$t_{.70}$	$t_{.60}$	$t_{.55}$
1	63.66	31.82	12.71	6.31	3.08	1.376	1.000	.727	.325	.158
2	9.92	6.96	4.30	2.92	1.89	1.061	.816	.617	.289	.142
3	5.84	4.54	3.18	2.35	1.64	.978	.765	.584	.277	.137
4	4.60	3.75	2.78	2.13	1.53	.941	.741	.569	.271	.134
5	4.03	3.36	2.57	2.02	1.48	.920	.727	.559	.267	.132
6	3.71	3.14	2.45	1.94	1.44	.906	.718	.553	.265	.131
7	3.50	3.00	2.36	1.90	1.42	.896	.711	.549	.263	.130
8	3.36	2.90	2.31	1.86	1.40	.889	.706	.546	.262	.130
9	3.25	2.82	2.26	1.83	1.38	.883	.703	.543	.261	.129
10	3.17	2.76	2.23	1.81	1.37	.879	.700	.542	.260	.129
11	3.11	2.72	2.20	1.80	1.36	.876	.697	.540	.260	.129
12	3.06	2.68	2.18	1.78	1.36	.873	.695	.539	.259	.128
13	3.01	2.65	2.16	1.77	1.35	.870	.694	.538	.259	.128
14	2.98	2.62	2.14	1.76	1.34	.868	.692	.537	.258	.128
15	2.95	2.60	2.13	1.75	1.34	.866	.691	.536	.258	.128
16	2.92	2.58	2.12	1.75	1.34	.865	.690	.535	.258	.128
17	2.90	2.57	2.11	1.74	1.33	.863	.689	.534	.257	.128
18	2.88	2.55	2.10	1.73	1.33	.862	.688	.534	.257	.127
19	2.86	2.54	2.09	1.73	1.33	.861	.688	.533	.257	.127
20	2.84	2.53	2.09	1.72	1.32	.860	.687	.533	.257	.127
21	2.83	2.52	2.08	1.72	1.32	.859	.686	.532	.257	.127
22	2.82	2.51	2.07	1.72	1.32	.858	.686	.532	.256	.127
23	2.81	2.50	2.07	1.71	1.32	.858	.685	.532	.256	.127
24	2.80	2.49	2.06	1.71	1.32	.857	.685	.531	.256	.127
25	2.79	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
26	2.78	2.48	2.06	1.71	1.32	.856	.684	.531	.256	.127
27	2.77	2.47	2.05	1.70	1.31	.855	.684	.531	.256	.127
28	2.76	2.47	2.05	1.70	1.31	.855	.683	.530	.256	.127
29	2.76	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
30	2.75	2.46	2.04	1.70	1.31	.854	.683	.530	.256	.127
40	2.70	2.42	2.02	1.68	1.30	.851	.681	.529	.255	.126
60	2.66	2.39	2.00	1.67	1.30	.848	.679	.527	.254	.126
120	2.62	2.36	1.98	1.66	1.29	.845	.677	.526	.254	.126
180	2.58	2.33	1.96	1.645	1.28	.842	.674	.524	.253	.126

PERCENTILE VALUES (χ_p^2) for THE CHI-SQUARE DISTRIBUTION

with ν degrees of freedom
(shaded area = p)

ν	$\chi^2_{.995}$	$\chi^2_{.99}$	$\chi^2_{.975}$	$\chi^2_{.95}$	$\chi^2_{.90}$	$\chi^2_{.75}$	$\chi^2_{.50}$	$\chi^2_{.25}$	$\chi^2_{.10}$	$\chi^2_{.05}$	$\chi^2_{.025}$	$\chi^2_{.01}$	$\chi^2_{.005}$
1	7.88	6.63	5.02	3.84	2.71	1.32	.455	.102	.0158	.0039	.0010	.0002	.0000
2	10.6	9.21	7.38	5.99	4.61	2.77	1.39	.575	.211	.103	.0506	.0201	.0100
3	12.8	11.3	9.35	7.81	6.25	4.11	2.37	1.21	.584	.352	.216	.115	.072
4	14.9	13.3	11.1	9.49	7.78	5.39	3.36	1.92	1.06	.711	.484	.297	.207
5	16.7	15.1	12.8	11.1	9.24	6.63	4.35	2.67	1.61	1.15	.831	.554	.412
6	18.5	16.8	14.4	12.6	10.6	7.84	5.35	3.45	2.20	1.64	1.24	.872	.676
7	20.3	18.5	16.0	14.1	12.0	9.04	6.35	4.25	2.83	2.17	1.69	1.24	.989
8	22.0	20.1	17.5	15.5	13.4	10.2	7.34	5.07	3.49	2.73	2.18	1.65	1.34
9	23.6	21.7	19.0	16.9	14.7	11.4	8.34	5.90	4.17	3.33	2.70	2.09	1.73
10	25.2	23.2	20.5	18.3	16.0	12.5	9.34	6.74	4.87	3.94	3.25	2.56	2.16
11	26.8	24.7	21.9	19.7	17.3	13.7	10.3	7.58	5.58	4.57	3.32	3.05	2.60
12	28.3	26.2	23.3	21.0	18.5	14.8	11.3	8.44	6.30	5.23	4.40	3.57	3.07
13	29.8	27.7	24.7	22.4	19.8	16.0	12.3	9.30	7.04	5.89	5.01	4.11	3.57
14	31.3	29.1	26.1	23.7	21.1	17.1	13.3	10.2	7.79	6.57	5.63	4.66	4.07
15	32.8	30.6	27.5	25.0	22.3	18.2	14.3	11.0	8.55	7.26	6.26	5.23	4.60
16	34.3	32.0	28.8	26.3	23.5	19.4	15.3	11.9	9.31	7.96	6.91	5.81	5.14
17	35.7	33.4	30.2	27.6	24.8	20.5	16.3	12.8	10.1	8.67	7.56	6.41	5.70
18	37.2	34.8	31.5	28.9	26.0	21.6	17.3	13.7	10.9	9.39	8.23	7.01	6.26
19	38.6	36.2	32.9	30.1	27.2	22.7	18.3	14.6	11.7	10.1	8.91	7.63	6.84
20	40.0	37.6	34.2	31.4	28.4	23.8	19.3	15.5	12.4	10.9	9.59	8.26	7.43
21	41.4	38.9	35.5	32.7	29.6	24.9	20.3	16.3	13.2	11.6	10.3	8.90	8.03
22	42.8	40.3	36.8	33.9	30.8	26.0	21.3	17.2	14.0	12.3	11.0	9.54	8.64
23	44.2	41.6	38.1	35.2	32.0	27.1	22.3	18.1	14.8	13.1	11.7	10.2	9.26
24	45.6	43.0	39.4	36.4	33.2	28.2	23.3	19.0	15.7	13.8	12.4	10.9	9.89
25	46.9	44.3	40.6	37.7	34.4	29.3	24.3	19.9	16.5	14.6	13.1	11.5	10.5
26	48.3	45.6	41.9	38.9	35.6	30.4	25.3	20.8	17.3	15.4	13.8	12.2	11.2
27	49.6	47.0	43.2	40.1	36.7	31.5	26.3	21.7	18.1	16.2	14.6	12.9	11.8
28	51.0	48.3	44.5	41.3	37.9	32.6	27.3	22.7	18.9	16.9	15.3	13.6	12.5
29	52.3	49.6	45.7	42.6	39.1	33.7	28.3	23.6	19.8	17.7	16.0	14.3	13.1
30	53.7	50.9	47.0	43.8	40.3	34.8	29.3	24.5	20.6	18.5	16.8	15.0	13.8
40	66.8	63.7	59.3	55.8	51.8	45.6	39.3	33.7	29.1	26.5	24.4	22.2	20.7
50	79.5	76.2	71.4	67.5	63.2	56.3	49.3	42.9	37.7	34.8	32.4	29.7	28.0
60	92.0	88.4	83.3	79.1	74.4	66.0	59.3	52.3	46.5	43.2	40.5	37.5	35.5
70	104.2	100.4	90.0	90.5	85.5	77.6	69.3	61.7	55.3	51.7	48.8	45.4	43.3
80	116.3	112.3	106.6	101.9	96.6	88.1	79.3	71.1	64.3	60.4	57.2	53.5	51.2
90	128.3	124.1	118.1	113.1	107.6	98.6	89.3	80.6	73.3	69.1	65.6	61.8	59.2
100	140.2	135.8	129.6	124.3	118.5	109.1	99.3	90.1	82.4	77.9	74.2	70.1	67.3

CONVERSION OF PEARSON'S r INTO CORRESPONDING FISHER'S z COEFFICIENT*

$$z = \frac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) = 1.513\log_{10}\left(\frac{1+r}{1-r}\right)$$

r	z	r	z	r	z	r	z	r	z	r	z
.25	.26	.40	.42	.55	.62	.70	.87	.85	1.26	.950	1.83
.26	.27	.41	.44	.56	.63	.71	.89	.86	1.29	.955	1.89
.27	.28	.42	.45	.57	.65	.72	.91	.87	1.33	.960	1.95
.28	.29	.43	.46	.58	.66	.73	.93	.88	1.38	.965	2.01
.29	.30	.44	.47	.59	.68	.74	.95	.89	1.42	.970	2.09
.30	.31	.45	.48	.60	.69	.75	.97	.90	1.47	.975	2.18
.31	.32	.46	.50	.61	.71	.76	1.00	.905	1.50	.980	2.30
.32	.33	.47	.51	.62	.73	.77	1.02	.910	1.53	.985	2.44
.33	.34	.48	.52	.63	.74	.78	1.05	.915	1.56	.990	2.65
.34	.35	.49	.54	.64	.76	.79	1.07	.920	1.59	.995	2.99
.35	.37	.50	.55	.65	.78	.80	1.10	.925	1.62		
.36	.38	.51	.56	.66	.79	.81	1.13	.930	1.66		
.37	.39	.52	.58	.67	.81	.82	1.16	.935	1.70		
.38	.40	.53	.59	.68	.83	.83	1.19	.940	1.74		
.39	.41	.54	.60	.69	.85	.84	1.22	.945	1.78		

*r's under 0.5, may be taken as equivalent to z's.

SPEARMAN'S RANK DIFFERENCE CORRELATION
(For one-tailed test)

N	.05	.01
5	.900	1.000
6	.829	.943
7	.714	.893
8	.643	.833
9	.600	.783
10	.564	.746
12	.506	.712
14	.456	.645
16	.425	.601
18	.399	.564
20	.377	.534
22	.359	.508
24	.343	.485
26	.329	.465
28	.317	.448
30	.306	.432

(For two-tailed test the P value of .05 and .01 will be .10 and .02 respectively)

AREAS OF STANDARD NORMAL DISTRIBUTION

An entry in the table is the proportion under the entire curve which is between $z = 0$ and a positive value of z. Areas for negative values of z are obtained by symmetry.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2703	.2734	.4764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4666	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4318
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4632
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4708
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

APPENDIX 3

Table value of *t* at different degree of freedom on P = 0.05 and .01 level.

P / V	.05	.01
1	6.314	31.821
2	2.920	6.965
3	2.353	4.541
4	2.132	3.747
5	2.015	3.365
6	1.943	3.143
7	1.895	2.998
8	1.860	2.896
9	1.833	2.821
10	1.812	2.764
11	1.796	2.718
12	1.782	2.681
13	1.771	2.650
14	1.761	2.624
15	1.753	2.602
16	1.746	2.583
17	1.740	2.567
18	1.734	2.552
19	1.729	2.541
20	1.725	2.528
21	1.721	2.518
22	1.717	2.508
23	1.714	2.500
24	1.711	2.492
25	1.708	2.485
26	1.706	2.479
27	1.703	2.463
28	1.701	2.467
29	1.699	2.462
30	1.697	2.457
40	1.684	2.423
60	1.671	2.390
- 120	1.658	2.338

APPENDIX 4

Distribution of χ^2.

Degree of freedom	*Probability*		
	0.05	*0.01*	*0.001*
1	3.84	6.64	10.63
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.29	18.47
5	11.07	15.09	29.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.99	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.69	29.14	36.12
15	25.00	30.58	37.70
16	26.30	32.00	39.25
17	27.59	33.41	40.79
18	28.87	34.81	42.31
19	30.14	36.19	43.82
20	31.41	37.57	45.32
21	32.67	38.93	46.80
22	33.92	40.29	48.27
23	35.17	41.64	49.73
24	36.42	42.98	51.18
25	37.65	44.31	52.62
26	38.89	45.64	54.05
27	40.11	46.96	55.48
28	41.34	48.28	56.89
29	42.56	49.59	58.30
30	43.77	50.89	59.70

APPENDIX 5

The correlation coefficient, r^2.

Degree of freedom	*Probability*		
	0.05	*0.01*	*0.001*
1	0.997	1.000	1.000
2	0.950	0.990	0.999
3	0.878	0.959	0.991
4	0.811	0.917	0.974
5	0.755	0.875	0.951
6	0.707	0.834	0.925
7	0.666	0.798	0.989
8	0.632	0.765	0.872
9	0.602	0.735	0.847
10	0.576	0.708	0.823
11	0.523	0.684	0.801
12	0.532	0.661	0.780
13	0.513	0.641	0.760
14	0.497	0.623	0.742
15	0.482	0.606	0.725
16	0.468	0.590	0.708
17	0.456	0.575	0.693
18	0.444	0.561	0.679
19	0.433	0.549	0.665
20	0.423	0.537	0.652
25	0.381	0.487	0.597
30	0.349	0.449	0.554
35	0.325	0.418	0.519
40	0.304	0.393	0.490
45	0.288	0.372	0.465
50	0.273	0.354	0.443
60	0.250	0.325	0.408
70	0.232	0.302	0.380
80	0.217	0.283	0.357
90	0.205	0.267	0.338
100	0.195	0.254	0.321

